

PHOTOGRAPH THIS SHEET

DTIC FILE COPY

INVENTORY

AD-A223 722

DTIC ACCESSION NUMBER

LEVEL

AFOSR 90 0696 APP. E
DOCUMENT IDENTIFICATION
1988

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

DISTRIBUTION STATEMENT

ACCESSION FOR

NTIS GRA&I ☒

DTIC TAB ☐

UNANNOUNCED ☐

JUSTIFICATION

BY

DISTRIBUTION /

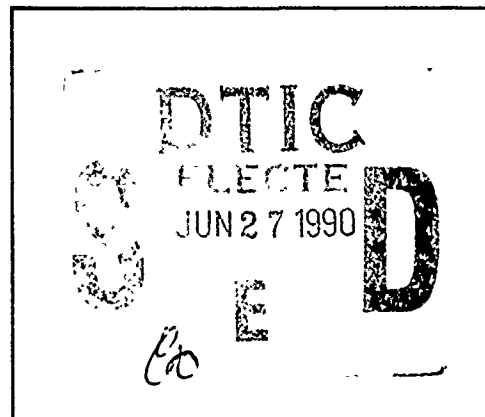
AVAILABILITY CODES

DIST

AVAIL AND/OR SPECIAL

A-1

DISTRIBUTION STAMP



DATE ACCESSIONED

DATE RETURNED

90 06 25 042

DATE RECEIVED IN DTIC

REGISTERED OR CERTIFIED NO.

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-PDAC

LOAN DOCUMENT

TRANSMITTAL

THE ATTACHED DOCUMENT HAS BEEN
LENT TO DTIC FOR PROCESSING. DO
NOT MARK OR MUTILATE THIS COPY.
PLEASE GIVE IT SPECIAL HANDLING
SO THE DOCUMENT MAY BE RETURNED
TO THE LENDER PROMPTLY . RETURN
THIS REPORT TO THE SELECTION
SECTION, DTIC-~~BD~~AC.

DO NOT PHOTOGRAPH THIS FORM
LOAN DOCUMENT

REPORT DOCUMENTATION PAGE

Form Approved
OASD No. 0706-0128

Public Reporting Burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0706-0128), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE	3. REPORT TYPE AND DATES COVERED Annual	
4. TITLE AND SUBTITLE Laboratory Graduate Fellowship Program <i>Appendix E</i>			5. FUNDING NUMBERS 61102F 23087D6 3484/06	
6. AUTHOR(S) Dr. Darrah, Lt. Col Claude Cavender				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Universal Energy Systems Inc.			8. PERFORMING ORGANIZATION REPORT NUMBER AEOSR-TR- 90 0696	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/XOT Bld 410 Bolling AFB, D.C. 20332-6448			10. SPONSORING / MONITORING AGENCY REPORT NUMBER F49620-86-C-0127	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 400 words) See Attached				
14. SUBJECT TERMS			15. NUMBER OF PAGES	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT N/A	

AD-A223 722

APPENDIX E

THESIS

I. INTRODUCTION

Critical to the success of the Air Force Office of Scientific Research (AFOSR) mission is the ability of AFOSR to draw upon the research community in the United States to respond to its needs. In recent years, however, the number of U. S. citizens seeking advanced degrees in the areas of Air Force research interests has been decreasing. This refers specifically to the number of U. S. citizens obtaining Ph.D. degrees in areas of mathematics and science that are of interest to the Air Force. This situation points toward the potential problem of a future shortage of qualified researchers in areas critical to the nation's security interest.

To address this problem, the United States Air Force Laboratory Graduate Fellowship Program (USAF/LGFP) was established. The contract is funded under the Air Force Systems Command by the AFOSR. The program annually provides three-year fellowships for at least 25 Ph.D. students in research areas of interest to the Air Force. Universal Energy Systems, Inc. (UES) has completed the third year of the three-year LGF program contract.

This report, prepared in compliance with contractual requirements, covers the third year of the program which now sponsors 27 first-year participants as well as 25 second-year fellows and 22 third year fellows for a total of 74 active fellowships. The report addresses an overview of the administration tasks, statistics on the 1989 awards, profiles of all the fellows, and summarized results of the evaluation process. Materials deemed inappropriate for inclusion in the main body of the report, such as samples of forms, complete questionnaire results, etc., are included in the appendices.

II. ADMINISTRATION

The administration of the LGF program is conducted from the Dayton offices of UES. The staff consists of Mr. Rodney C. Darrah, Program Manager; Ms. Judy Conover, Program Administrator; and support personnel. Most members of the 1989 program administration team have been involved with the project since award of the contract to UES. This element of an experienced, stable staff ensures program continuity and contributes to successful operation of administrative tasks.

The primary tasks in managing the program consist of advertising (which includes compiling and updating a mailing list, and preparing and distributing ads, flyers, and

DISSERTATION

ELECTROMAGNETIC EXCITATIONS ON ANTIFERROMAGNETS:
SURFACE POLARITONS AND LEAKY WAVES

Submitted by

Robert L. Stamps

Physics Department

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 1988

COLORADO STATE UNIVERSITY

Fall 1988

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED
UNDER OUR SUPERVISION BY ROBERT L. STAMPS ENTITLED
ELECTROMAGNETIC SURFACE EXCITATIONS ON ANTIFERROMAGNETS BE
ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

R. E. Comley

Robert L. Stamps

W. J. ...

Chiang-yun Shen

Adviser

Robert L. Stamps
Department Head

ABSTRACT OF DISSERTATION
ELECTROMAGNETIC EXCITATIONS ON ANTIFERROMAGNETS:
SURFACE POLARITONS AND LEAKY WAVES

Surface polaritons exist at frequencies between 250 GHz and a few THz when propagating on antiferromagnets. They have been predicted theoretically and observed recently on MnF_2 in reflection measurements. Although damping is necessary for coupling between the surface polariton modes and incident light waves, theoretical studies have not examined the effect of damping on the surface polariton modes. An analysis of the surface polariton modes, with damping, is done both by solving the dispersion equation and by studying the classical electromagnetic Green's functions for a semi-infinite antiferromagnet. In addition to modifying the surface polariton dispersion curves, damping also allows for the existence of new surface modes in regions otherwise forbidden to surface wave propagation. These new modes are analogous to the Brewster and evanescent modes found for electromagnetic surface waves on metals and dielectrics. The Green's functions are also applied to the problem of scattering light off of a rough surface antiferromagnet and are used to show that surface roughness allows the incident wave to excite surface polaritons and leaky modes. Surface roughness is found to enhance nonreciprocal reflectance (a reflectance which is different for incidence at angle θ and $-\theta$). It has been argued that nonreciprocal reflection cannot exist without nonreciprocal absorption processes in the material.

Depending on the polarization of the incident wave, however, it is shown that nonreciprocal reflection *can* exist without absorption. In an example calculation, the reflectance of an antiferromagnet is found to be highly nonreciprocal when the the illuminating beam is circularly polarized.

Robert L. Stamps
Physics Department
Colorado State University
Fort Collins, CO 80531
Fall 1988

ACKNOWLEDGMENTS

I gratefully acknowledge the Physics Department of the University of Colorado at Colorado Springs, where much of this work was done, for generously allowing me the use of their facilities. I also want to thank the Physics Department of Colorado State University for allowing me the freedom to work closely with the University of Colorado. I am especially grateful to Professor Carl E. Patton for his assistance and support, and to the members of his magnetism group whose assistance and knowledge made my time at Colorado State enjoyable and rewarding. Most especially, I want to thank Professor R. E. Camley, University of Colorado, for the opportunity to participate in these projects. I thank him for his patience, his guidance, and his unwavering encouragement and enthusiasm.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	iii
ACKNOWLEDGMENTS	v
TABLE OF CONTENTS	vi
<u>CHAPTER</u>	
1 INTRODUCTION	1
2 MAGNETIC EXCITATIONS AND POLARITONS	6
2.1 Magnetic Interactions	6
2.2 Antiferromagnetic Susceptibility	12
2.3 Magnetostatic Waves	13
2.4 Bulk Antiferromagnetic Polaritons	15
2.5 Surface Antiferromagnetic Polaritons	21
2.6 Nonreciprocity	23
2.7 Damping Effects	26
2.8 Surface Roughness	30
3 NONRECIPROCAL REFLECTION	34
3.1 Possibility of nonreciprocal reflectance	34
3.2 Theory	41
3.3 Results	47

4	POLARITONS WITH DAMPING	61
	4.1 Surface Polaritons With Bloch-Bloembergen Damping	61
	4.2 Surface Polaritons With Landau-Lifshitz Damping	78
5	ELECTROMAGNETIC GREEN'S FUNCTIONS	83
	5.1 Green's Functions	84
	5.2 Surface Polaritons	96
6	SCATTERING FROM ROUGH SURFACES	101
	6.1 Born Approximation For Scattered Fields	103
	6.2 One Dimensional Grating	117
	6.3 Randomly Rough Surface	131
7	CONCLUSIONS	138
	7.1 Bulk Polaritons and Nonreciprocal Reflection	139
	7.2 Leaky Surface Modes On Antiferromagnets	139
	7.3 Scattering From Periodic Gratings	140
	7.4 Scattering From Rough Surfaces	141
	7.5 Future Extensions	141
	REFERENCES	143
	APPENDICES	145
	A Fresnel Relations	145
	B Green's Functions	147
	C Landau-Lifshitz Damping Terms	150

LIST OF TABLES

<u>Table</u>		<u>Page</u>
I	Physical Parameters For Antiferromagnets	11
II	Characteristics of Leaky Modes	29

LIST OF FIGURES

<u>Figure</u>		<u>Page</u>
2.1	The Antiferromagnetic Lattice	7
2.2	Bulk And Surface Spin Waves	9
2.3	Dispersion Curves for $H_0=0$	18
2.4	Dispersion Curves for $H_0=.3\text{kG}$	19
2.5	Precession Modes	20
2.6	Symmetry Argument	25
2.7	Reduced Brillouin Zone For Grating Induced Scattering	33
3.1	Reflection Geometry	36
3.2	Detailed Balance Argument	37
3.3	Reflectance and Transmittance for Linear Polarizations	49
3.4	Change in Reflectance for Left Circular Polarizations	50
3.5	Change in Reflectance for Right Circular Polarizations	54
3.6	Change in Reflectance for Different ϕ	56
3.7	Change in Reflectance in Different Bulk Bands	57
3.8	Reflectance and Transmittance for $\omega/\Omega=1.006$	59

4.1	Dispersion Curves with Damping for $H_0=0$	65
4.2	Brewster Angle and Leaky Mode Dispersion Curve	67
4.3	Decay Parameters for Leaky Modes, $1/\Omega\tau=.0002$	68
4.4	Decay Parameters for Leaky Modes, $1/\Omega\tau=.0008$	70
4.5	Dispersion Curves with Damping for $H_0=.3kG$	71
4.6	Decay Parameters for Leaky Modes with $H_0=.3kG$	73
4.7	Power Flows in the Material and Vacuum for $H_0=0$	75
4.8	Power Flows in the Material and Vacuum for $H_0=.3kG$	77
4.9	Dispersion Curves with Landau Damping	82
5.1	Poles of g_{xx}	97
5.2	Dispersion Curve and g_{xx}	100
6.1	Rough Surface Geometry	104
6.2	Dispersion Curves with Grating Lines, $H_0=0$	121
6.3	Power Flows on Grating with $1/\Omega\tau=.0002$	123
6.4	Power Flows on Grating with $1/\Omega\tau=.0008$	125
6.5	Dispersion Curves with Grating Lines, $H_0=.3kG$	127
6.6	Evanescent Power Flows for the Grating	128
6.7	Reflectance of the Grating	130
6.8	Evanescent Power Flows for a Random Surface, $\sigma=.1$	134
6.9	Evanescent Power Flows for a Random Surface, $\sigma=.5$	135
6.10	Reflectance of the Rough Surface	137

CHAPTER 1

INTRODUCTION

The coupling between electromagnetic waves and the fundamental excitations in a material (plasmons, phonons, etc.) produces what is known as a polariton. Polaritons in dielectrics and metals have been extensively studied both theoretically and experimentally. Surface polaritons in dielectrics and metals, where the amplitude of the excitation is confined to the region near the surface, have also received a great deal of attention¹ and are now used as a tool in studying the vibrational spectra of very thin films.²

In contrast, magnetic polaritons³ (coupled electromagnetic waves and spin waves) and magnetic surface polaritons⁴ have received less attention. One reason for this is that in ferromagnets the frequency of the magnetic polariton is typically lower (1-20 GHz for metallic ferromagnets) than the frequency for a phonon-polariton. This means wavelengths for ferromagnetic polaritons are in the millimeter and centimeter range and are much larger than phonon-polaritons wavelengths. It is thus difficult in ferromagnets to realistically obtain an effectively infinite or semi-infinite sample size or to obtain information about the region close to the surface.

Recently, the properties of bulk and surface polaritons on a semi-infinite uniaxial *antiferromagnet* were discussed theoretically and measured by Remer, et al, on MnF_2 using reflectivity measurements.⁵ These excitations possess several intriguing features. First, and in contrast to the ferromagnet, antiferromagnetic polaritons have frequencies in the infrared with typical values ranging from 250 GHz to a few THz. The wavelengths (and penetration depths) thus range from millimeters to a few hundred micrometers. Second, and in contrast to phonon-polaritons, antiferromagnetic surface polaritons are nonreciprocal,⁶ i.e.

$\omega(+k) \neq \omega(-k)$ where ω is the frequency of the polariton and k is the wavevector. For nonreciprocal surface waves, reversing the direction of propagation or reversing the direction of an applied magnetic field generally leads to a surface polariton of a different frequency. The degree of nonreciprocity can be controlled by varying the strength of the applied field and disappears when the applied field is turned off.

The nonreciprocity of ferromagnetic and antiferromagnetic surface polaritons is a primary reason for studying these excitations. From a theoretical point of view, the existence or nonexistence of reciprocity is an interesting problem in its own right. The possibility for nonreciprocal features of an excitation in a complex system can often be arrived at by considering certain symmetries of the system, and so a knowledge of the conditions under which nonreciprocity exists helps to determine the key interactions which govern the excitations. As an example, an examination of the Hamiltonian for a ferromagnet gives the interesting result that it is not exchange interactions but rather the dipole-dipole interactions and the presence of a boundary on the material that determine nonreciprocity in ferromagnetic magnons.⁷

Knowledge of nonreciprocal features are clearly important for experimental investigations. The measurements by Remer of surface antiferromagnetic polaritons, for example, relied on the nonreciprocity of the surface polaritons in an applied magnetic field to provide a reflectance which was nonreciprocal with respect to the direction of the applied field. Other examples where nonreciprocity can easily be exploited are magnetic superlattices and antiferromagnetic thin films. These systems are expected to support nonreciprocal spin waves even without applied magnetic fields. In the superlattice, the nonreciprocity depends simply on whether there is an even or odd number of layers. In a similar manner, the nonreciprocity in a thin film with (100) surfaces depends on whether there is an even or odd number of (100) sublattice planes across the film.

Nonreciprocal properties also have a number of important technological applications.^{8,9} A list of existing devices used in microwave signal processing which

depend on nonreciprocal features include circulators, isolators, phase shifters, delay lines, resonators and signal to noise enhancers. At microwave frequencies, these devices depend on the nonreciprocity of ferromagnetic magnetostatic modes and are commonly manufactured out of YIG (yttrium iron garnet). A circulator allows the multiple connection of different devices (such as connecting a receiver and transmitter to the same antenna) by using the property that ferromagnetic magnetostatic waves on thick films only propagate in one direction. An isolator uses a nonreciprocal localization of waves in a film to one surface or the other in order to separate signals. Phase shifters and delay lines use applied fields to control the group velocity of a magnetostatic wave propagating on a ferrite surface.

Any nonreciprocal feature is potentially interesting from an applications point of view. By analogy to the great many uses of nonreciprocal ferromagnetic devices in microwave signal processing devices, one expects there to be a corresponding number of applications for antiferromagnetic devices at infrared frequencies. In fact, recent proposals for nonreciprocal magnetoplasmon based devices, to operate in the 300 to 600 GHz range, indicate an interest for devices at high frequencies.¹⁰

As discussed above, surface polaritons on antiferromagnets have already been observed and their basic features catalogued and understood. In terms of discussing the nonreciprocal properties of antiferromagnetic polaritons, however, only the nonreciprocal frequency behavior $\omega(k) \neq \omega(-k)$ of the surface modes has been examined. The bulk polariton modes, on the other hand, are reciprocal in frequency in applied magnetic fields. Their polarization, however, depends on the magnitude and direction of the external field. This dissertation begins with an examination of the coupling between photons and bulk antiferromagnetic polariton modes in external magnetic fields and shows that there can be a nonreciprocal reflectance of an incident wave such that the $R(\theta) \neq R(-\theta)$ where θ is the angle of incidence.

The existing studies of surface polaritons on antiferromagnets have essentially considered only a very idealized situation where the antiferromagnetic film is perfect and

there is no scattering of the wave either due to imperfections in the crystal or interactions with other modes. The next step toward understanding the surface polaritons and assessing their potential utility is to ask how imperfections affect the surface polaritons.

Two types of imperfections are considered. The first is irregularities intrinsic to the material that cause the surface wave to lose energy as it propagates. These irregularities are loosely termed "damping mechanisms". The second irregularity to be considered is extrinsic: geometrical imperfections on the surface of the material that can scatter the surface wave as it propagates. Both of these irregularities are important on two counts: first because damping and surface roughness are always present in reality to some degree and must be accounted for in either experiment or application; second, and perhaps most importantly, damping and roughness represent two distinct possible mechanisms for coupling the polariton fields to external photon fields. This is clearly important in any optical measurement or application of antiferromagnetic materials. With this in mind, the consequences of material damping on the antiferromagnetic surface electromagnetic excitations are examined. Next, the excitations are studied by using surface electromagnetic response functions for the material. Finally, the response functions are used to estimate the reflectance of a rough surfaced antiferromagnet.

A surprising result of including damping into the antiferromagnet and searching for possible surface excitations is the existence of "leaky" surface modes. These are excitations that depend on damping for their existence and have finite lifetimes and path lengths. They are localized to the surface, but transmit energy into the bulk of the material. An interesting feature is that the leaky modes exist in frequency and wavelength regions where true surface modes cannot. Leaky modes have been predicted and observed for non-magnetic materials and the leaky surface modes found in this magnetic system are analogous to the Brewster surface polariton modes found in dielectric materials.

The outline of this dissertation is as follows. Chapter 2 is an overview of electromagnetic excitations in antiferromagnets and gives a qualitative discussion of the effects of including damping and roughness into the description of the surface polaritons. In

Chapter 3 the reflectance of light from an idealized semi-infinite antiferromagnet is calculated and the nonreciprocity of the reflectance with respect to the incident angle is explained. In Chapter 4 the effects of damping are explored by examining the surface polariton dispersion curves obtained from Maxwell's equations and the properties of the leaky modes are examined in detail. The surface electromagnetic response functions, or Green's functions, are derived in Chapter 5 for a semi-infinite antiferromagnet and applied to the problem of scattering light from a rough surface in Chapter 6. Finally, Chapter 7 summarizes the main findings of this work and comments on possible future extensions.

CHAPTER 2

MAGNETIC EXCITATIONS AND POLARITONS

Polaritons are excitations in the electromagnetic field created by a variety of interacting systems of particles. A gas of weakly interacting electrons, for example, is a medium in which plasmon polaritons can exist. If a magnetic field is present, then magnetoplasmon polaritons are possible. The electromagnetic excitations in magnetic systems, such as ferromagnets and antiferromagnets, are known as magnon polaritons. To understand a particular polariton excitation, it is first necessary to understand the electrical and magnetic interactions within the system that support the excitation.

2.1 Magnetic interactions.

It is useful to develop a phenomenological model of the antiferromagnetic crystal lattice and describe the properties of this system in terms of macroscopic fields. This approach is valid as long as the excitations have long wavelengths relative to the lattice site spacing. In this limit, the electric and magnetic fields are averaged over distances of several lattice sites and the actual fluctuating microscopic fields are replaced by macroscopic effective fields.

Imagine a simple cubic lattice with magnetic moments placed at all the lattice sites. Suppose further that the moments are reversed from site to site, forming lattice planes with alternating magnetizations. There are then two sublattices forming the system, and the magnetization at any point in the crystal is the sum of an average magnetization \vec{M}_a due to one sublattice and an average magnetization \vec{M}_b due to the other sublattice. Such a system is pictured in figure 2.1 and can be used as a model for the uniaxial antiferromagnet MnF_2 .

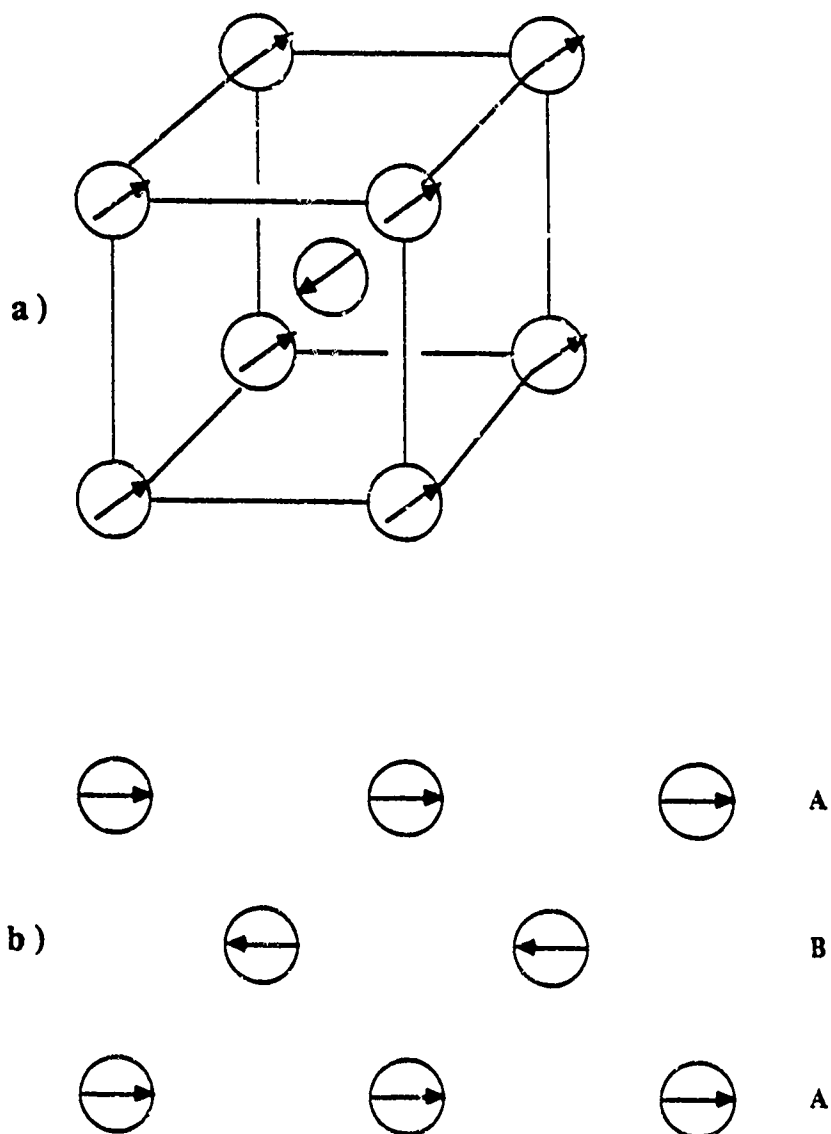


Figure 2.1. The antiferromagnetic crystal is represented by a periodic arrangement of microscopic magnetic moments M_a and M_b . All the A sublattice moments point in the same direction and all the B sublattice moments point in the opposite direction. The cubic structure shown represents MnF_2 .

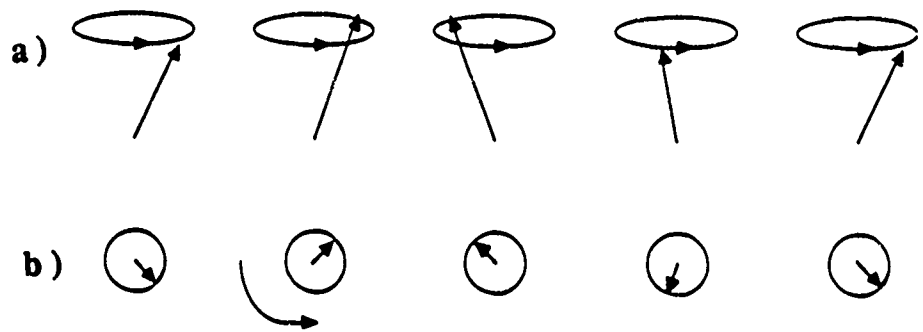
The magnitude of the two magnetizations are equal in an antiferromagnet and they point in opposite directions so that the total magnetization of the crystal is zero. Here only uniaxial antiferromagnets are considered where at equilibrium \vec{M}_a and \vec{M}_b are aligned along one particular direction in the crystal when there are no external magnetic fields. This direction is called the easy axis. A magnetic excitation is represented by small deviations of the moments at each site on the sublattices away from the easy axis. The deviations are coupled through the interacting moments and can be represented as waves which propagate through the crystal. This is illustrated in figure 2.2 where the deviations are due to the precession of magnetic moments about a magnetic field \vec{H}_{eff} . The force law governing the motion of the magnetizations is known as Bloch's equation of motion:

$$\dot{\vec{M}} = \gamma \vec{M} \times \vec{H}_{\text{eff}} \quad (2.1)$$

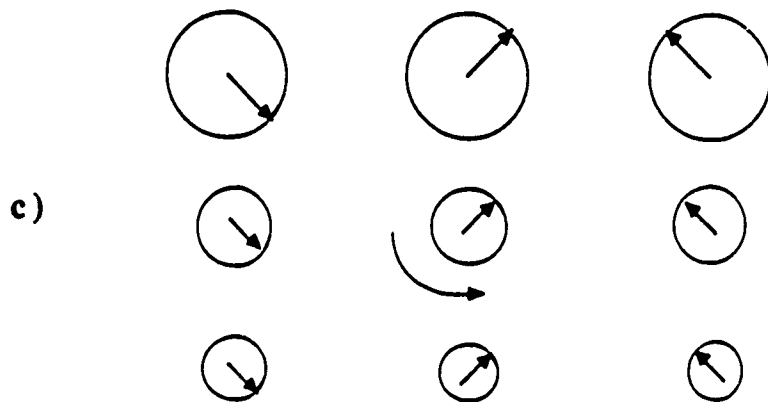
The gyromagnetic ratio is γ and has the value $-1.805 \times 10^{10} \text{ rads/kG}$.

The interactions between the individual magnetic moments are, on a macroscopic level, treated as effective fields acting on each sublattice. One is thus able to define two effective fields \vec{H}_{eff}^a and \vec{H}_{eff}^b and write two coupled equations of motion of the Bloch form. In the limit of small deviations from the equilibrium directions, these equations can be cast in the form $\dot{\vec{M}} = \chi \vec{h}$, where \vec{h} is the macroscopic dipolar field, and used to find an explicit form for the magnetic susceptibility χ . An explicit example of this calculation is given in chapter 4 where the susceptibilities are derived when Landau damping is included into the Bloch equations of motion.

The effective fields have several parts: first, there is a static externally applied field \vec{H}_0 , which will always be aligned along the easy axis. Anisotropy fields, which act to direct the moments along their respective sublattices, are defined by the constant H_a . The macroscopic dipolar fields, h , are defined by $\vec{B} = \vec{h} + 4\pi\vec{M}$. Finally, the average effect of the exchange interaction, which couples each moment to its nearest neighbors, is represented



Bulk mode



Surface mode

Figure 2.2. (a) The precession of the magnetizations about the effective fields is shown by the circles which represent the path traced out by the tip of the magnetization vector. (b) An end view of the precessions where the effective field is normal to the plane of view. (c) The precession angles of the magnetizations decrease with distance into the material for waves localized to the surface.

through a macroscopic field \vec{H}_{ex} in the mean field limit. This limit assumes that the field varies slowly from lattice site to lattice site where the mean field is directly proportional to the average sublattice magnetizations. This restricts the description to frequencies where the wavelength of the excitation spans several lattice sites. Demagnetization fields are not taken into account for two reasons. First, the geometry used throughout this dissertation is semi-infinite with the surface plane parallel to the easy axis. The net magnetization normal to the easy axis is zero and so there are no demagnetization fields in directions normal to the surface. The net magnetization is zero parallel to the surface and so there are also no demagnetization fields in these directions either.

With these definitions, the effective field acting on sublattice A is given by

$$\vec{H}_{eff}^a = \hat{z}(H_0 + H_a) + \lambda \vec{M}_b + \vec{h} \quad (2.2a)$$

The quantity $\lambda \vec{M}_b$ is the mean exchange field acting on sublattice A due to moments on sublattice B. λ is an exchange constant that can be obtained from a microscopic exchange Hamiltonian. Note that the easy axis coincides with the z axis in this definition.

The effective field acting on sublattice B is given by the similar expression

$$\vec{H}_{eff}^b = \hat{z}(H_0 - H_a) + \lambda \vec{M}_a + \vec{h} \quad (2.2b)$$

Here, $\lambda \vec{M}_a$ is the mean exchange field acting on sublattice B due to moments on sublattice A. In both (2.1) and (2.2), the anisotropy fields are along the easy axis. These fields direct the moments on the A sublattice in the +z direction and the moments on the B sublattice in the -z direction. Values of these parameters are listed in Table I for the uniaxial antiferromagnets MnF_2 ,¹¹ $GdAlO_3$,¹² and FeF_2 .¹³ These materials are well characterized and have been studied with microwave, Brillouin scattering, and far infrared techniques.

Table I. Physical parameters of example uniaxial antiferromagnets.

	MnF ₂	GdAlO ₃	FeF ₂
Neel Temperature (K)	67	3.87	79
Exchange field H _{ex} (kG)	550	18.8	540
Anisotropy field H _a (kG)	7.87	3.65	200
Sublattice magnetization M(kG)	.6	.624	.56
Resonance field Ω(kG)	93.37	12.26	506

2.2 Antiferromagnetic Susceptibility.

The magnetic susceptibility $\vec{\chi}$ is derived from Bloch's equations using (2.1) and (2.2) and assuming the time dependence of the magnetizations and the dipolar fields varies as $e^{-i\omega t}$. Small deviations in the x and y directions are assumed and only quantities of first order in M_x, M_y and h_x, h_y are retained.

A useful quantity for Maxwell's electromagnetic equations is the permeability, defined by $\vec{\mu} = \vec{1} + 4\pi\vec{\chi}$. For the uniaxial antiferromagnet, this has the form

$$\vec{\mu} = \begin{bmatrix} \mu_1 & i\mu_2 & 0 \\ -i\mu_2 & \mu_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.3)$$

Gaussian units are assumed throughout this dissertation. The components of the tensor are given by :

$$\mu_1 = 1 - \frac{4\pi\gamma^2 M H_a}{\Omega^2} (X+Y) \quad (2.4)$$

$$\mu_2 = \frac{4\pi\gamma^2 M H_a}{\Omega^2} (Y-X) \quad (2.5)$$

Here M is the magnitude of the sublattice magnetizations and Ω is the antiferromagnetic resonance field given by

$$\Omega^2 = H_a \gamma^2 (H_a + 2H_e) \quad (2.6)$$

The quantities X and Y are defined as

$$X = [(\omega / \Omega + H_0 \gamma / \Omega + i / \Omega \tau)^2 - 1]^{-1} \quad (2.7)$$

and

$$Y = [(\omega / \Omega - H_0 \gamma / \Omega + i / \Omega \tau)^2 - 1]^{-1} \quad (2.8)$$

τ is a phenomenological spin relaxation time and is included as the simplest possible representation of a loss mechanism. It originates from the inclusion of the Bloch-Bloembergen damping term $-\tau \dot{\vec{M}}$ into the right hand side of the equations of motion (2.1).¹⁴

The the magnetic behavior of the antiferromagnet is described with the frequency dependant $\vec{\mu}$. The material is dissipative when τ is finite and has a resonance at frequency Ω . At resonance, the moments on each sublattice precess about the easy axis uniformly in phase with one another.

2.3 Magnetostatic Waves.

When the phase of the precessing moments on each sublattice varies in a periodic manner, as in figure 2, the excitations are called spin waves. When the wavelengths of these excitations are very long, they exist at frequencies that satisfy the condition $k \gg \omega/c$. Then $\nabla \times \vec{H} = 0$ and the waves are independent of any electric fields. These excitations are called magnetostatic spin waves. In planar ferromagnetic structures they are often called Damon Eshbach modes and in spherical ferromagnetic structures, they are known as Walker modes.

There are many intriguing and useful properties of these waves. In both ferromagnets and antiferromagnets, the magnetostatic spin waves exist only in certain frequency regions, or bands. In very thick films the bulk modes can propagate at the frequencies¹⁵

$$\omega_1 = \gamma [H_a(2H_e + H_a)]^{1/2}$$

and

$$\omega_2 = \gamma [H_a(2H_e + H_a + 8\pi M)]^{1/2}$$

Note that ω_1 is also the antiferromagnetic resonance frequency.

In thin film structures, the bulk modes over a range of frequencies, or bands, but only certain wavevectors are allowed at any frequency within a bulk band. Also, applied fields drastically affect the allowed frequencies and wavevectors of the modes. In antiferromagnets, new frequency bands appear with the application of an applied field.

A particularly intriguing feature is the existence of surface magnetostatic waves; i.e., waves which are localized to the surface of a structure. These waves propagate along the surface, decaying exponentially away in directions normal to the surface. In the magnetostatic region, the surface modes exist in frequency regions forbidden to bulk waves. In very thick films, surface modes lie between the bulk bands:

$$\omega_s = \gamma [H_a(2H_e + H_a + 4\pi M)]^{1/2}$$

Surface modes are also strongly affected by applied magnetic fields. The presence of external fields leads to antiferromagnetic magnetostatic surface spin waves that have different frequencies for opposite propagation directions. This is an example of nonreciprocity in frequency with respect to propagation direction, a property discussed more fully in connection with antiferromagnetic polaritons in chapter 3.

2.4 Bulk Antiferromagnetic Polaritons.

Electromagnetic excitations also occur at wavelengths where $k > \omega/c$ and the fluctuating magnetic fields have associated electric fields. These electromagnetic excitations are polaritons and exist at frequencies near those of magnetostatic waves but have much longer wavelengths. Just as photons are the excitations of the electromagnetic field in vacuum, these polaritons are excitations of the electromagnetic field in the material. The behavior of these excitations is governed by the magnetic susceptibilities (2.8) and also by the dielectric susceptibilities. The constitutive relations $\vec{B} = \mu \vec{H}$ and $\vec{D} = \epsilon \vec{E}$ are used which obey the retarded Maxwell equations:

$$\nabla \cdot \vec{D} = 0$$

$$\nabla \cdot \vec{B} = 0$$

$$\nabla \times \vec{E} + \frac{1}{c} \frac{\partial \vec{B}}{\partial t} = 0$$

$$\nabla \times \vec{H} - \frac{1}{c} \frac{\partial \vec{D}}{\partial t} = 0$$

Materials which are anisotropic in their dielectric properties will be considered here. The dielectric susceptibilities are written in tensoral form as

$$\vec{\epsilon} = \begin{bmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_2 \end{bmatrix}$$

(2.9)

The dielectric susceptibilities are assumed to be constant over the frequency ranges of interest here, although it is a simple matter to substitute the appropriate frequency dependent functions for ϵ_1 and ϵ_2 .

It is a straight forward exercise to write the electromagnetic wave equation from the Maxwell relations. In vector form, it is given by

$$\nabla \times \overleftrightarrow{\epsilon}^{-1} \nabla \times \vec{H} - \omega_0^2 \overleftrightarrow{\mu} \vec{H} = 0 \quad (2.10)$$

Here and throughout the rest of this dissertation, $\omega_0 = \omega/c$. Note that the magnetic fields \vec{H} are assumed to have the time dependence $e^{-i\omega t}$. In the less obvious, but computationally more efficient form, this is equivalent to

$$\epsilon_k \sum_{m,j} ' \left\{ \epsilon_m^{-1} \frac{\partial^2}{\partial x_j \partial x_k} - \delta_{jk} \left(\sum_l \epsilon_m^{-1} \frac{\partial^2}{\partial x_l^2} \right) - \omega_0^2 \mu_{kj} \right\} H_j(\vec{x}) = 0 \quad (2.11)$$

The prime on the sum means that $m \neq j, k, l$. Also, since the dielectric tensor is diagonal, $\epsilon_{kk} = \epsilon_k$.

For the present, assume the material extends infinitely in all directions. The solutions to this equation then represent bulk antiferromagnetic polaritons in an infinite or semi-infinite media. Plane wave solutions are assumed with the form $e^{i\vec{k} \cdot \vec{x} - i\omega t}$ where \vec{k} is the wavevector of the excitation and ω is the frequency.

Using plane wave solutions for \vec{H} , an expression for $\vec{k}(\omega)$ can be obtained from the wave equation (2.11). For arbitrary directions of propagation, $\vec{k}(\omega)$ is rather complicated. For the special case of propagation in a direction perpendicular to the easy axis, the dispersion relation is particularly simple and has the form¹⁶

$$k^2 = k_x^2 + k_y^2 = \omega_0^2 \epsilon_2 \left(\frac{\mu_1^2 - \mu_2^2}{\mu_1} \right) \quad (2.12)$$

Here k_x and k_y are the x and y components of the wavevector k , respectively.

First consider the case where there is no applied field. Then μ_2 is identically zero and the frequency dependence of the wavevector is determined by μ_1 . In frequency regions where μ_1 is positive, k is real and the polaritons are travelling waves. In regions where μ_1 is negative, k is imaginary and the polaritons decay exponentially.

Since μ_1 has a pole at the antiferromagnetic resonance frequency, it becomes infinitely large as ω approaches the resonance frequency and also changes sign. Likewise, k becomes infinite at this frequency. For $\mu_1 > 0$, k is real but for $\mu_1 < 0$, k is imaginary. The frequency regions where k is real are the bulk polariton bands. It is only in these regions that the antiferromagnet is transparent to electromagnetic radiation.

In figure 2.3, these bulk bands are depicted in the dispersion curve ω vs. k_x . The quantities shown are in reduced units, ω/Ω and ck_x/Ω , a convention used throughout this dissertation. The material is MnF_2 , there is no applied field and $\omega/\Omega=1$ is the resonance frequency. The bulk bands are represented by the shaded areas and represent regions where k_y (as a function of ω and k_x) is real. There are two bulk bands, one bounded above by the resonance frequency and extending indefinitely to lower frequencies, and one bounded below and extending indefinitely to higher frequencies.

In an applied field, μ_2 is no longer zero and a third bulk band appears. This is depicted in figure 2.4 where an applied field of .3kG is applied in the z direction. The top of the middle bulk band corresponds to the pole of $1/\mu_1$ which occurs at ω_2 , the high frequency magnetostatic bulk mode described in section 2.3. In zero applied field, this frequency represents the bottom limit of the high frequency bulk polariton band shown in figure 2.3. In both figures 2.3 and 2.4, k_y is imaginary outside the bulk bands.

Some insight into the nature of the bulk modes can be had by examining the resonance motion of the magnetizations \vec{M}_a and \vec{M}_b . In zero applied field, these two vectors can precess about the easy axis in two different ways.¹⁷ In one case, the motion is

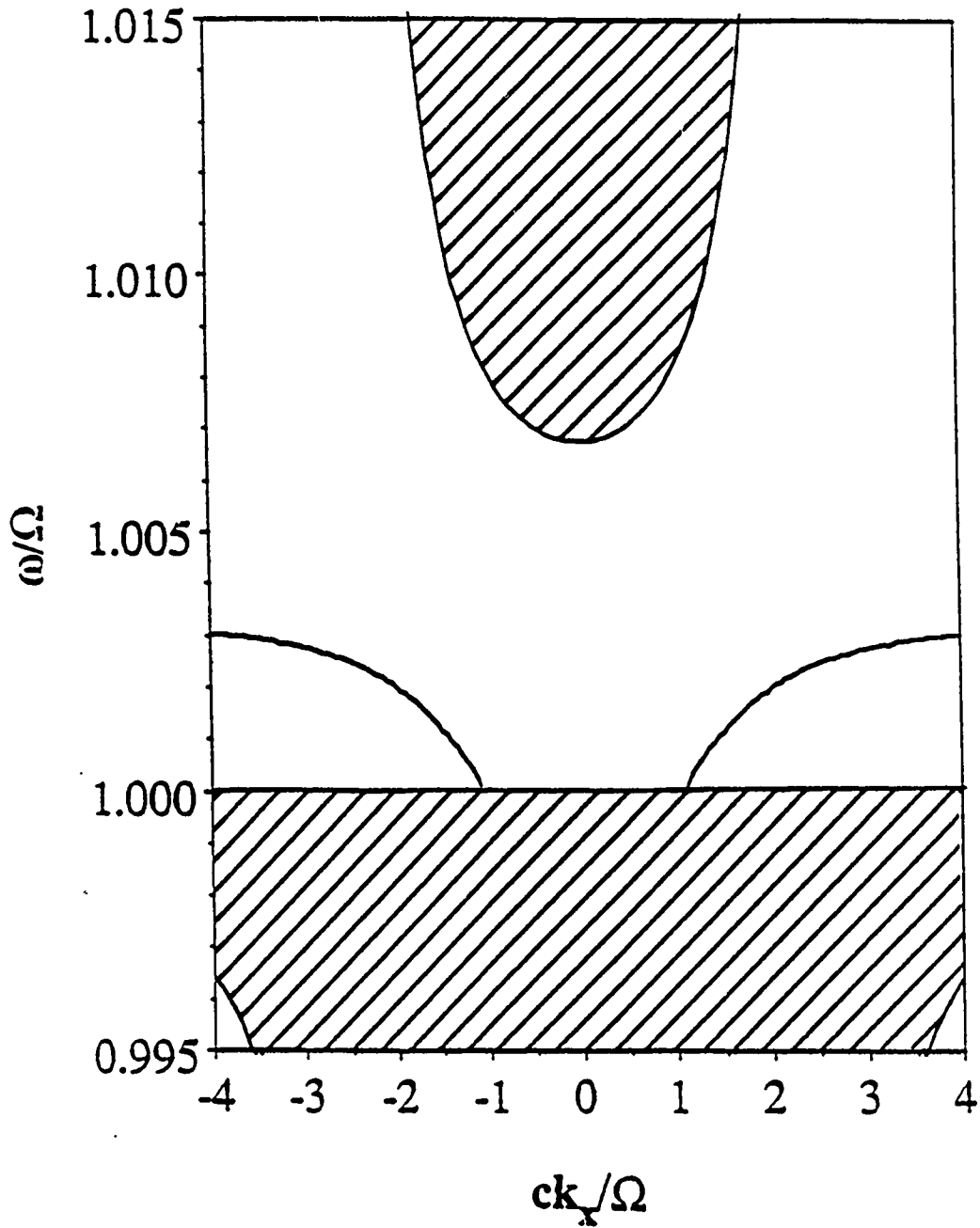


Figure 2.3. Dispersion curves for antiferromagnetic polaritons, in MnF_2 , with no applied field. The shaded areas are bulk bands and the solid lines rising out of the bulk bands are the surface modes when there is no damping present.

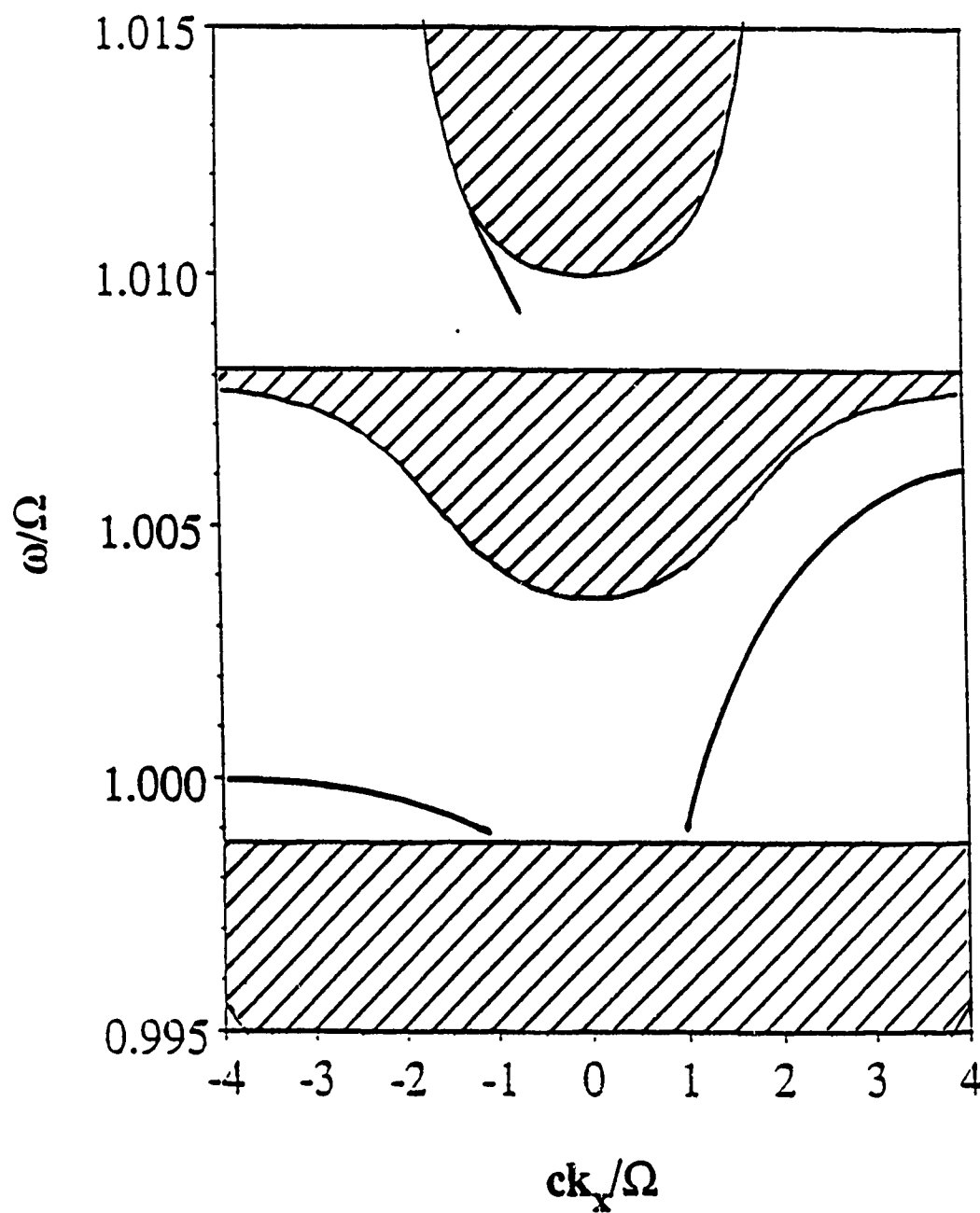


Figure 2.4. Dispersion curves for antiferromagnetic polaritons, in MnF_2 , with an applied field of .3kG. The shaded areas are bulk bands and the solid lines rising out of the bulk bands are the surface modes when there is no damping present in the material.

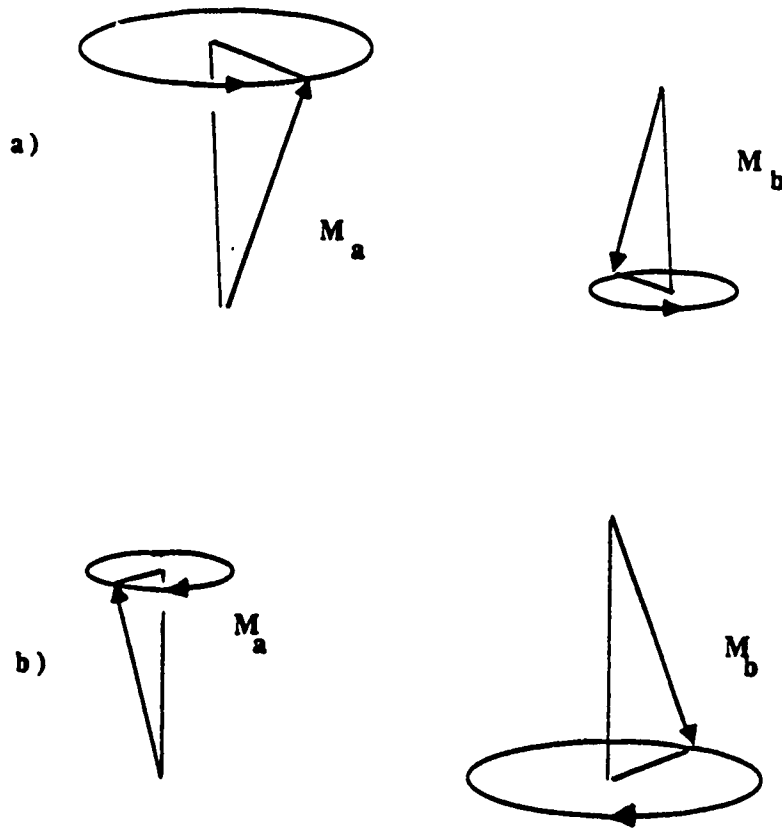


Figure 2.5. Here the precession of the two sublattice magnetizations are compared in zero applied field. Each magnetization precesses about the effective field acting on its sublattice. In zero applied field, there are two degenerate possible rotations: one in which the A sublattice magnetization precesses with a larger angle than the B sublattice magnetization and one in which the B sublattice magnetization precesses with a larger angle than the A sublattice magnetization. The two motions differ in their sense of rotation. In an applied field the two rotations still have opposite directions but are no longer degenerate.

clockwise and the precession angle for \vec{M}_a is larger than that of \vec{M}_b . In the other case, the motion is counter clockwise and the precession angle for \vec{M}_a is smaller than that of \vec{M}_b . These two motions are depicted in figure 2.5.

In zero applied field, the energies of these two precessions are degenerate. With an applied field, however, the two motions are no longer equivalent and have different frequencies. Since the bulk polaritons are coupled to the precession of the magnetizations, they should somehow also be affected by the applied field. It turns out that the polarization of the bulk polaritons is controlled by the applied field and this will be discussed in detail in the next chapter.

2.5 Surface Antiferromagnetic Polaritons.

When the structure is finite so that an interface exists between the antiferromagnet and some dissimilar media, a new form of polariton can exist that is localized to the interface.

These are called surface polaritons and differ from the bulk polaritons in many significant respects. Instead of a continuum of frequencies at a given wavelength, as in bulk modes, surface polaritons typically exist at only a finite number of frequencies. They may also display a high degree of nonreciprocity with respect to frequency; i.e., $\omega(\vec{k}) \neq \omega(-\vec{k})$.

Surface polaritons are characterized by an exponential decrease in amplitude away from the interface. In this dissertation, this interface will always be between the antiferromagnet and vacuum. If the normal to the surface is in the y direction, as shown in figure 2.6, exponential decay corresponds to an imaginary k_y both in the antiferromagnet and outside it. Inside the material, k_y is imaginary when μ_1 is negative. This occurs for frequencies just above the resonance pole. Outside the material in vacuum, k_y is imaginary only for frequencies and wavevectors away from the light line $\omega/k = c$ such that $\omega/k < c$. This means surface polaritons exist outside frequency and wavelength regions accessible to bulk polaritons on either side of the interface.

A dispersion relation for the surface polaritons can be obtained by imposing boundary conditions at the surface. For the electromagnetic waves considered here, these boundary conditions are the usual continuity conditions on tangential \vec{H} and \vec{E} fields and normal \vec{D} and \vec{B} fields. The procedure is to apply the boundary conditions to the bulk solutions of the wave equation (2.11) and the bulk solutions of the wave equation appropriate to free space across the surface. The solutions in both regions are required to decay exponentially away from the surface and the boundary conditions result in a characteristic equation that can be solved for the relationship between ω and k .

For surface waves travelling in the x direction according to wavevector k_x , the solutions inside and outside the material ($y>0$ and $y<0$, respectively) thus have the form

$$A e^{i\vec{k} \cdot \vec{x} - \alpha y} \quad y > 0 \quad (2.13)$$

$$B e^{i\vec{k} \cdot \vec{x} + \gamma y} \quad y < 0 \quad (2.14)$$

where α is the decay parameter inside the material (always assumed positive) and is given by (2.12) with $k_y = i\alpha$. Outside the material, γ is the decay parameter (also always assumed positive) and is given by $[k_x^2 - \omega_0^2]^{1/2}$. Continuity of tangential H and normal B lead to the implicit dispersion relation¹⁶

$$\gamma + (\mu_1 \alpha + \mu_2 k_x) / (\mu_1^2 - \mu_2^2) = 0 \quad (2.15)$$

The solutions to this equation are shown in figure 2.3 as the solid lines rising out of the lower bulk band and asymptotically approach the magnetostatic frequency ω_s as ω/k tends to zero. The existence of two surface modes, one for $+k_x$ and one for $-k_x$

(corresponding to propagation in the + and - x directions), is characteristic of antiferromagnetic surface excitations. Ferromagnetic magnetostatic surface modes for example, have only one branch. Note that both branches of the antiferromagnetic surface polariton dispersion curve begin at $\omega=ck$, the light line.

In an applied field, the character of the two modes changes radically. In figure 2.4 the solid lines again represent surface polariton modes, but this time the $+k_x$ and $-k_x$ branches have quite different shapes and exhibit the nonreciprocity of the surface modes in applied fields. Both of these branches begin at the light line and approach the magnetostatic frequencies. Also, there is an interesting new surface polariton branch above the middle bulk band which begins at the light line and approaches the outer edge of the upper bulk band.

2.6 Nonreciprocity.

The nonreciprocity of frequency with respect to wavevector, as in figure 2.4, is one of the most striking features of the antiferromagnetic surface polaritons. That it should exist is somewhat surprising since short wavelength spin waves do not exhibit this nonreciprocity. The origin of the nonreciprocity, though not completely understood, is a phenomena of the long wavelength excitations and is due to the dipolar fields and the presence of a surface.

It is possible to construct arguments that show the possibility of nonreciprocal behavior without requiring its existence. One such argument, originally due to Scott and Mills, is based on simple symmetry considerations. Consider first a material extending infinitely in all directions. A magnetic field is applied in the +z direction and a wave propagates in the $+k_x$ direction. Now look for some set of symmetry operations that take k_x to $-k_x$ while leaving the material unchanged with respect to the applied field.

A reflection in the yz plane changes k_x to $-k_x$ and, since the magnetic field is an axial vector, it also changes to point along the $-z$ direction. Another reflection, this time in the xz plane, returns the field to its original orientation but leaves k_x pointing in the $-x$ direction. These operations have reversed the wavevector and have returned the system to its original state. The equality $\omega(k_x) = \omega(-k_x)$ must hold for the bulk polaritons. A similar argument shows that $\omega(k_z) = \omega(-k_z)$.

Now suppose a semi-infinite material occupies the region $y > 0$ as in figure 2.6. Here the same sequence of operations fail to return the material to its original orientation by leaving the material in the upper half space. It is, in fact, impossible to take k_x to $-k_x$ and leave the material in place. There is thus no requirement that $\omega(k_x) = \omega(-k_x)$ for excitations propagating in the semi-infinite geometry. This argument shows the possibility for nonreciprocal behavior of a surface wave parallel to the surface, but it does not demand it. Nonreciprocity needs to be investigated case by case.

Nonreciprocity of frequency with respect to wavevector is only one possibility for nonreciprocal behavior. Another possibility for nonreciprocal behavior is in the reflection of light from magnetic materials at frequencies where the material is transparent. In this phenomena, the reflectance due to a wave incident at some angle of incidence θ would not equal the reflectance due to a wave incident at $-\theta$. Any nonreciprocal reflectance would be due to the energy transmitted into the material. In this case it is nonreciprocal behavior of the bulk polaritons which are of interest.

The above symmetry arguments ignore the polarization of the bulk polariton. In the next chapter the reflectance of an antiferromagnet in an applied field is examined and it is found that while the energies of the bulk polaritons are reciprocal with respect to propagation direction, the polarizations of the $+k_z$ and $-k_z$ modes can lead to a transmittance for an incident wave with $+k_z$ that differs from the transmittance for an incident wave with $-k_z$.

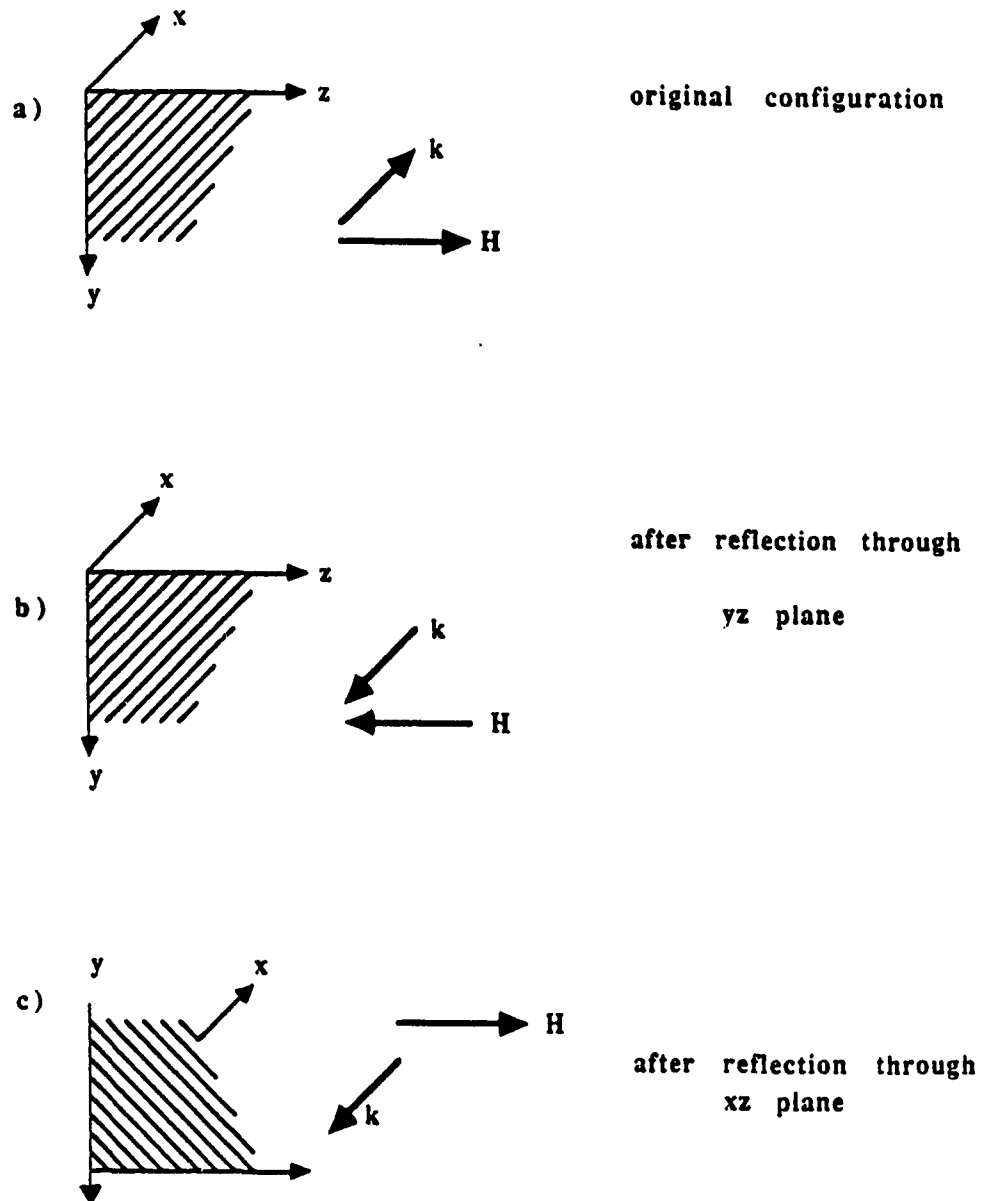


Figure 2.6. In (a) the initial configuration of the semi-infinite antiferromagnet is defined with the wavevector k of the excitation in the x direction, the material in the $y > 0$ half space, and the applied field H_0 directed into the paper. In (b), the system has been reflected through the yz plane, taking k to $-k$ and H_0 to $-H_0$. In (c) the system has been reflected through the xz plane, taking H_0 back to its initial direction into the paper and leaving k in the $-x$ direction. These operations, however, fail to return the crystal to its original state.

2.7 Damping effects.

Up to now, only bulk and surface polaritons in non-dissipative media ($\tau=\infty$ in the equations of motion (2.7 and 2.8)) have been considered. This has been appropriate for enumerating the properties of these excitations as true eigenmodes of the magnetic system and has simplified the analysis by allowing k_y to be either pure real or pure imaginary.

By ignoring dissipation, a number of important physical processes are not taken into account. These are primarily scattering events between magnons and magnons, magnons and phonons, etc., and can be thought of in a gross way as mechanisms of transferring energy out of one spin wave mode into other spin wave modes and other excitations in the crystal lattice. The simplest way to represent these energy losses is by introducing a finite lifetime τ into the equations of motion. The advantage of this representation is that it is mathematically simple to deal with. It is limited, however, by describing only relaxation processes that align the precessing dipoles along their equilibrium directions. This accounts for interactions between the spin system and the crystal lattice, but does not include interactions between spins that would tend to align the dipoles to local fields. A slightly more detailed description will be presented in chapter 3 when a damping term is introduced into the equations of motion that describes the relaxation of the spins to the direction of the instantaneous effective fields.

A finite τ leads to complex susceptibilities. The real part of the susceptibilities governs electromagnetic wave propagation in the material and the imaginary part determines the absorption of electromagnetic energy by the material. For the bulk and surface polaritons described above, the complex susceptibilities result in complex wavevectors. These complex wavevectors have a simple interpretation: the real part corresponds to propagating wave-like excitations, and the imaginary part represents excitations that decay exponentially as they travel.

One consequence of material damping is a slight modification of the bulk and surface polariton frequencies. The surface mode frequencies shown in figure 2.3, for example, are shifted slightly up and the bulk mode boundaries are also affected to a small degree. These effects are due to the shifting of the resonance pole in the susceptibilities.

By far the most drastic effect of damping is the existence of new polariton modes in frequency regions otherwise forbidden to polariton excitations. This phenomena has not been studied in antiferromagnetic systems before, but is very well known from plasmon-polariton studies. These modes are surface modes in the sense that they are bound to the surface of the material, but they can also have components that travel into the bulk of the material. These components are attenuated as they move away from the surface due to dissipation into the crystal. For this reason, they are also often referred to as "leaky modes" since they leak electromagnetic energy travelling along the surface into the interior of the crystal where it is eventually absorbed through the various mechanisms discussed above.

There is a wealth of terminology describing the leaky modes that exist on the interface between two dielectric media. The surface polariton mode, which exists even in the absence of damping, is termed the Fano mode. Modes which depend on damping for their existence fall into three classes: Evanescent, Brewster, and Zenneck modes. Evanescent modes are characterized as rapidly attenuating as they propagate, with a path length on the order of their wavelength. They are thus difficult to observe. They also are unique in that they exist in frequency regions forbidden to both surface and bulk polaritons. Zenneck modes exist in frequency regions where the dielectric function is almost completely imaginary.

The properties of leaky modes in dielectric materials are listed in Table II in terms of the relative magnitudes of the real and imaginary parts of the dielectric function and the wavevector parallel and perpendicular to the surface.¹ The real parts of the wavevector describe the propagation of the wave and the imaginary parts determine the attenuation of the wave. For example, the imaginary part of the surface polariton's parallel wavevector

component is much smaller than the real part. Surface polaritons thus have very long pathlengths. The perpendicular component, however, has an imaginary part that is much larger than the real part. Thus surface polaritons are tightly bound to the surface of the material. By comparison, the Brewster modes also have long pathlengths but are weakly bound to the surface of the material.

Some of the leaky surface modes found in the magnetic system are analogous to the Brewster surface polariton modes found in dielectric materials. As is well known, for a wave incident on a non-absorbing dielectric and E field polarized in the plane of incidence, there exists an angle of incidence for which there is no reflected wave. Since the Brewster angle depends on the frequency ω , it provides a relationship between ω and k_x , the component of the wavevector parallel to the surface. With damping present in the dielectric, one can find a solution of Maxwell's equations for a weakly bound surface wave which decays as it propagates parallel to the surface. This surface mode has a dispersion relation ω as a function of $\text{Re}(k_x)$ which is very similar to the relationship in the Brewster case and is known as a Brewster mode. In the antiferromagnetic system, a very equivalent result can be found. Here a Brewster angle occurs for a geometry where the magnetic field lies in the plane of incidence. Again when damping is present, a magnetic Brewster surface mode can exist with a dispersion relation similar to that provided by the relationship of the magnetic Brewster's angle and the frequency.

In chapter 4, a detailed discussion of the antiferromagnetic leaky modes and their properties is presented. Since these modes depend on damping for their existence, it is natural to question the form of the damping terms included into the equations of motion. In equations (2.7) and (2.8), only the simplest damping terms, Bloch-Bloembergen, are represented. This kind of damping describes a relaxation of the precessing magnetizations to their equilibrium values.

Table II. Properties of leaky waves in dielectrics. The dielectric function is ϵ and the wavevector parallel to the surface is k_x . The wavevector component normal to the surface is denoted by q where $q=k_m$ inside the material and $q=k_v$ in the vacuum.

Type	ϵ	k_x	q	Phase velocity
"True" Surface Polariton	$\text{Re}(\epsilon) < -1$ $\text{Im}(\epsilon) < < \text{Re}(\epsilon)$	$\text{Re}(k_x) \gg \text{Im}(k_x)$	$\text{Re}(q) < < \text{Im}(q)$	$< c$
Evanescent	$-1 < \text{Re}(\epsilon) < 0$ $\text{Im}(\epsilon) \approx \text{Re}(\epsilon) $	$\text{Re}(k_x) \approx \text{Im}(k_x)$	$\text{Re}(q) \approx \text{Im}(q)$	
Brewster	$\text{Re}(\epsilon) > 0$ $\text{Im}(\epsilon) < < \text{Re}(\epsilon)$	$\text{Re}(k_x) \gg \text{Im}(k_x)$	$\text{Re}(q) \gg \text{Im}(q)$	$> c$
Zenneck	$\text{Re}(\epsilon) < < -1$ $\text{Im}(\epsilon) \gg \text{Re}(\epsilon) $	$\text{Re}(k_x) \gg \text{Im}(k_x)$	$\text{Re}(q) \approx \text{Im}(q)$	$> c$

Alternatively, one can describe a relaxation to the instantaneous value of the effective fields. The required expressions can be derived from the free energy and are called Landau-Lifshitz damping terms. These terms are calculated and included in the susceptibilities in Appendix I. Their effect on the antiferromagnetic leaky modes is explored in chapter 4.

2.8 Surface Roughness.

Having examined the nature of magnetic excitations in the antiferromagnet, described the coupled photon and magnon modes that exist in the electromagnetic field of the crystal, and considered the effects of damping mechanisms on the surface electromagnetic excitations, this survey of surface and bulk polaritons is nearly complete. Throughout this investigation, the focus has been on quantities important for the experimental observation and potential utility of these excitations. Continuing in this manner, the following discussion considers the effect of surface roughness on polariton excitations and the possibility of coupling light to these excitations through periodic gratings ruled on the antiferromagnetic surface.

Surface roughness provides two mechanisms by which surface polaritons can lose energy: either through roughness induced radiation or through roughness induced scattering into different polariton states. Which mechanism dominates depends on the polariton's frequency. Thus surface roughness can affect the mean free path length of polaritons propagating on the surface of a material. These mechanisms have in fact been suggested as explanations of certain anomalies in attenuated total reflection measurements of plasmon polaritons.²⁵ Another effect due to random roughness is the shifting of frequencies of surface polariton states below their flat surface frequencies, thus displacing the dips in reflectivity curves by amounts dependant on the distribution of heights and profiles that characterize the roughness. Again, this is also a typical consequence of dissipation mechanisms.

One of the most promising aspects of surface roughness effects is the ability to couple incident light with surface excitations. Surface polaritons have phase velocities $\omega/k < c$ and so cannot directly couple with incident light. Damping in the material allows one possible mechanism for coupling by giving a width in frequency to the susceptibility poles. Surface modes with frequencies and wavevectors very near the light line can then be driven in an "off-resonance" manner.

Surface roughness creates a width in wavelength analogous to the width in frequency produced by damping. A very simple type of roughness, a periodic grating, clearly illustrates this. As is well known, light incident on a grating is diffracted into several directions according to the wavelength of the light and period of the grating. Each direction corresponds to a diffracted beam with a different parallel wavevector component. The parallel wavevector component of the scattered light is given by $k_x = k_{x0} \pm nk_g$ where k_{x0} is the parallel wavevector component of the incident wave, k_g is the grating period and n is an integer. The grating thus scatters the incident wave into an infinite number of parallel wavevector states that differ by integral multiples of the grating's spatial period. Many of these wavevector states k_x satisfy $k_x < \omega/c$ and so correspond to radiative states. States that have $k_x > \omega/c$, however, decay exponentially away from the surface and are evanescent waves. Energy in these states can couple directly with surface polariton modes and thus provide a means of transferring energy from the incident light wave to surface electromagnetic excitations.

This process is analogous to the scattering of electrons in periodic potentials, and one can construct Brillouin zone representations for the surface polariton modes on periodic gratings.^{18,19} This is shown schematically in figure 2.7 where grating induced "light lines" are superposed on the dispersion curve for a surface polariton. Each light line corresponds to one of the n diffracted beams and the incident beam can then couple to a surface polariton everywhere a light line intersects the polariton dispersion curve.

Other features reminiscent of energy bands is the appearance of "band gaps" in the surface polariton dispersion relations that exist at the Brillouin zone boundaries. Also, it is possible that the fields very near a rough surface are enhanced to such a degree as to significantly affect nonlinear optical processes.²⁰

By allowing an incident beam to couple with the surface polariton modes, surface roughness can *enhance* the nonreciprocal reflectivity of an antiferromagnet in an applied field. In other words, in an applied field the difference in reflectivity between a beam incident at $+\theta$ and a beam incident at $-\theta$, both of which couple to the surface polariton modes, is increased by the presence of roughness.

Since leaky modes also exist in regions away from the light line, roughness can allow coupling between an incident beam and the Brewster-like modes of a damped antiferromagnet. A theory for reflection from a rough antiferromagnetic surface is presented in chapter 5 with emphasis on coupling to surface excitations. Numerical results are presented for reflection from a sinusoidal grating and a randomly rough surface.

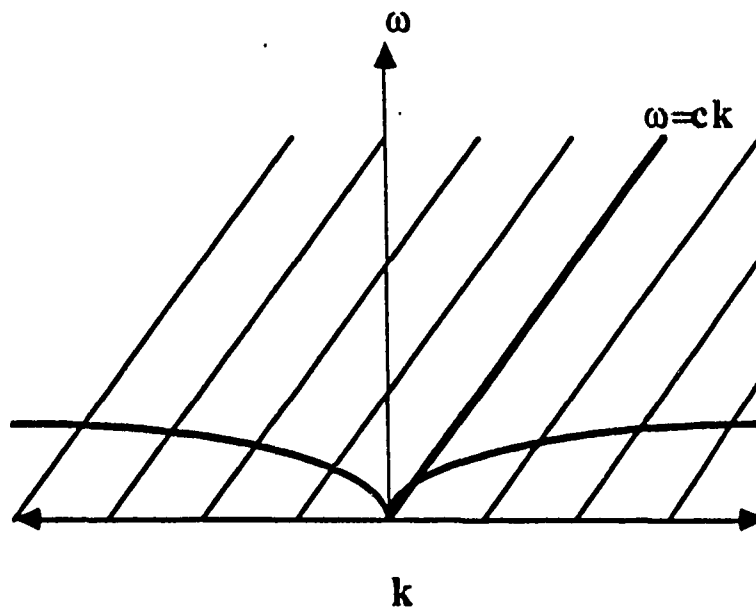


Figure 2.7. In analogy to energy bands for electrons in semiconductors, an extended zone scheme can be used to show the possible excitations of surface polariton modes. Shown here is a typical dispersion curve for surface polariton modes plotted as frequency vs. wavenumber. The solid straight line is the dispersion curve for light in vacuum where $\omega=ck$. Without the grating, only those surface modes which cross the light line $\omega=ck$ can be excited. With a grating, the light line is split into an infinite number of parallel lines, each with slope c , spaced a distance nk_g apart. The grating then allows the incident light to excite surface modes anywhere a grating line intersects the surface polariton curve. A similar figure results for light incident at negative incident angles except that all the light lines then have negative slope.

CHAPTER 3

NONRECIPROCAL REFLECTION

Nonreciprocal reflection, where the reflectance of a material for some angle of incidence θ differs from the reflectance at $-\theta$, has been observed on antiferromagnets by coupling the surface polaritons to an incident electromagnetic wave and measuring the resulting change in reflectance.⁵ This actually constituted the first experimental observation of antiferromagnetic surface polaritons. The coupling was done by means of damping mechanisms in the material which allowed energy to dissipate from the electromagnetic wave into the surface polariton modes. An interesting question, which is the topic of this chapter, is whether nonreciprocal reflection can exist without absorption present in the material. It turns out that nonreciprocal reflection with respect to incident angle can exist without the presence of material damping by coupling the incident electromagnetic wave directly with the bulk polariton modes.

3.1 Possibility of nonreciprocal reflectance.

Detailed balance arguments provide one means of showing the possibility (or lack thereof) for nonreciprocal reflection. One such an argument, originally due to Remer et al., shows that nonreciprocity in reflection is not possible without absorption in the material.²¹ First define the reflectance, $R(\theta, \phi, \omega, T)$, as the power of an electromagnetic wave with frequency between ω and $\omega+d\omega$ which is reflected by a surface element dF at a temperature T . This is normalized by the power normally incident on the surface element dF in a solid angle in the direction defined by (θ, ϕ) with frequency between ω and $\omega+d\omega$. The angle θ is the angle the radiation makes with respect to the normal of the surface, while ϕ is the angle

the reflection plane makes with the magnetic field (see figure 3.1). The direction (θ, ϕ) always refers to the incident wave. The transmittance $T(\theta, \phi, \omega, T)$ and absorptance $A(\theta, \phi, \omega, T)$ are defined similarly.

Now imagine a blackbody and the material interchanging radiation at temperature T . At equilibrium, the conservation of energy and detailed balance result in the equality

$$A(\theta) + R(\theta) = A(-\theta) + R(-\theta) = 1 \quad (3.1)$$

It would thus seem that nonreciprocal reflection is possible only in the presence of nonreciprocal absorption; that is, $R(\theta) - R(-\theta) \neq 0$ if $A(\theta) \neq A(-\theta)$. It has been pointed out, however, that this is true only for the net incident and reflected power, but not necessarily for the individual polarizations of the incident and reflected waves.⁷

The argument can be restated in more general form by defining separate reflectances, absorptances and transmittances for each polarization and considering sums over all polarization states. In the following, the polarization state of the incident energy is denoted by the subscript i and the polarization state of the corresponding reflected or transmitted energy is denoted by the subscript e . Also, the absorptance in each state, A_{ie} , is assumed to be zero.

Consider an infinite nonabsorbing slab surrounded on both sides by infinite blackbodies. Suppose both blackbodies and the slab are at equilibrium with uniform temperature T . Each blackbody radiates energy $I_{ie}(\theta)$ in the direction θ with polarization i . For this geometry, it is necessary to define separate reflectances and transmittances for each surface. These are labelled T_{1ie} , T_{2ie} , R_{1ie} , and R_{2ie} as illustrated in figure 3.2(a). The numeric subscript on any quantity identifies which surface the corresponding incident energy illuminated.

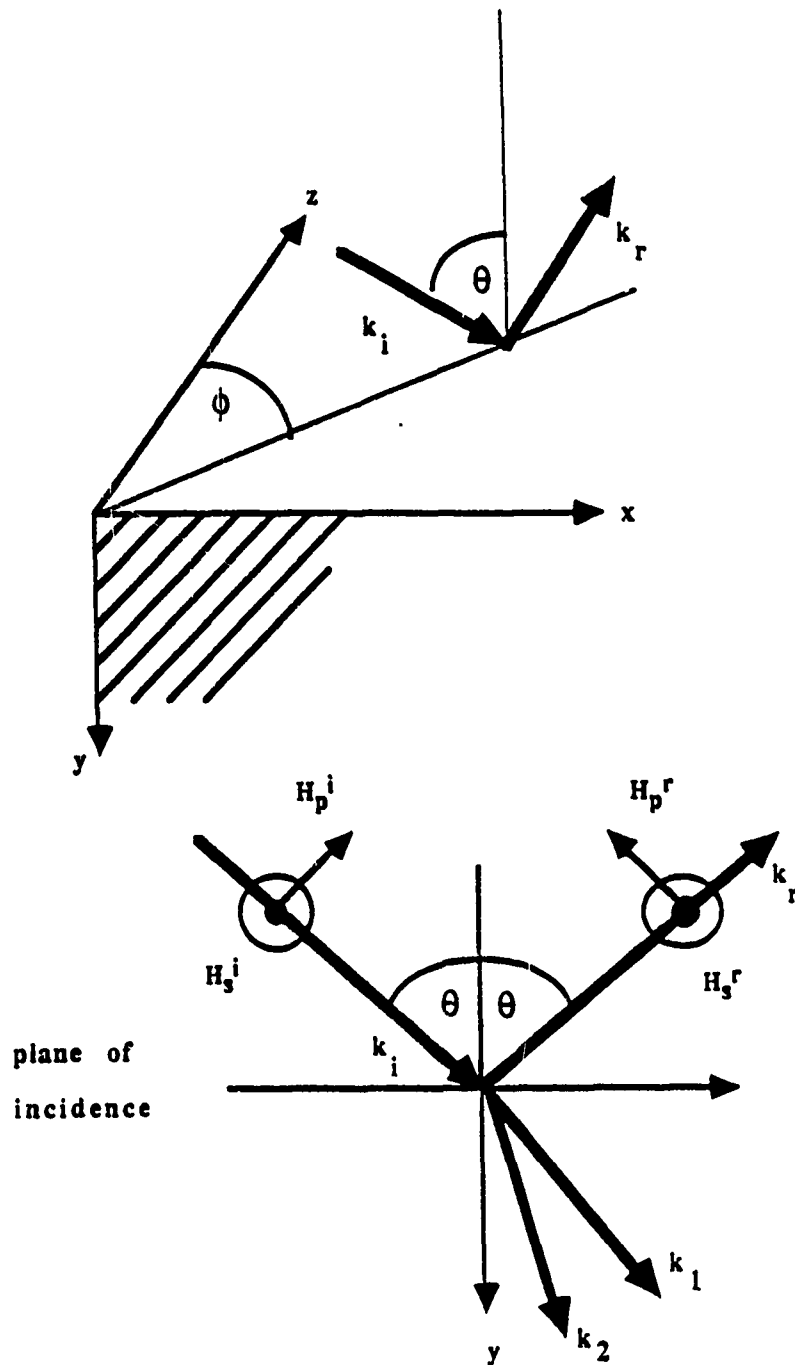


Figure 3.1. Reflection geometry. The material occupies the upper half plane, $y > 0$, and the surface is in the xz plane. The saturation magnetizations and applied field lie along the z axis. The angle of incidence, θ , is the angle the wavevector of the incident wave makes with the normal to the surface. The plane of incidence makes an angle ϕ with respect to the z axis. The incident magnetic fields H_s^i and H_p^i represent the components of the incident magnetic field perpendicular and parallel to the plane of incidence. The reflected magnetic fields H_s^r and H_p^r are defined similarly. Note that for $\theta \neq 90^\circ$, there are two transmitted waves which are denoted by the wavevectors k_1 and k_2 with magnetic fields H_1 and H_2 .

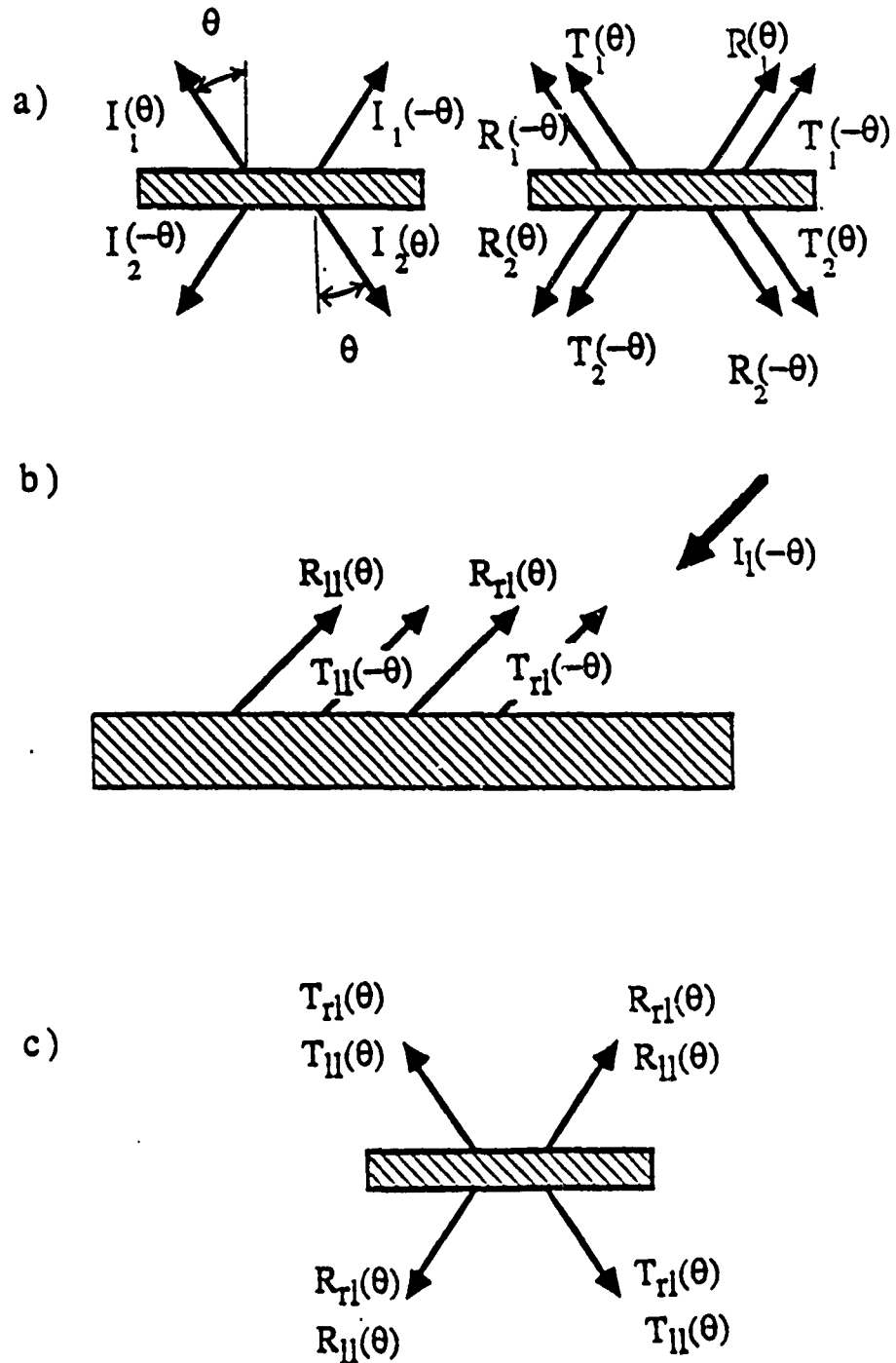


Figure 3.2. The net transmitted energy emitted by the slab from surface 1 is T_1 and the net energy reflected by surface 1 is R_1 . T_2 and R_2 are defined similarly. The angles θ and $-\theta$ refer to the angle of incidence.

$T_{1ie}(\theta)$ and $R_{1ie}(\theta)$ represent the transmitted and reflected energy, with polarization e , due to the energy incident on surface 1 at an angle θ with polarization i . $T_{2ie}(\theta)$ and $R_{2ie}(\theta)$ are the transmitted and reflected energy with polarization e , due to the energy incident on surface 2 at an angle θ with polarization i .

In order to make use of the symmetry of this problem, the energy incident in the $+\theta$ direction on surface 1 illuminates the slab from the left. The corresponding $R_{1ie}(\theta)$ then leaves the surface toward the right. Similarly, the energy incident in the $+\theta$ direction on surface 2 is defined as coming from the right. The corresponding $R_{2ie}(\theta)$ then leaves the surface toward the left. Thermal equilibrium requires the energy incident on surface 1 from the upper blackbody to equal the energy incident on surface 2 from lower blackbody. With these definitions, $R_{1ie}(\theta) = R_{2ie}(\theta)$ and likewise for the transmittances. The subscripts identifying the surface are no longer necessary and are neglected for simplicity.

The blackbodies radiate energy in all polarization states equally. For equilibrium to exist in all polarization states, detailed balance requires the sum over all *incident* polarization states of the energy emitted with the polarization state e into the direction θ equal the energy in the same polarization state e incident in the direction θ . The energy emitted into the direction θ is composed of reflected energy (from energy incident in the direction $-\theta$) and transmitted energy (from energy incident from the direction θ). Thus:

$$I_e(\theta) = \sum_i (R_e(-\theta) + T_{ie}(\theta)) \quad (3.2)$$

Similarly, the energy emitted into the direction $-\theta$ (reflection of energy incident in the direction θ and transmission of energy incident in the direction $-\theta$) with polarization e summed over all incident polarization states must equal the energy with polarization e incident in the direction $-\theta$. The energy incident at $-\theta$ with polarization e equals the energy incident at $+\theta$ with polarization e and the following equality must hold:

$$\sum_i (R_{ie}(\theta) + T_{ie}(-\theta)) = \sum_i (R_{ie}(-\theta) + T_{ie}(\theta)) \quad (3.3)$$

This equality is shown schematically in figure 3.2(b) for the polarization states r and l.

A second equality can also be obtained. In figure 3.2(c) the energy emitted by the slab, due to energy with polarization l incident in the direction $+\theta$, is shown. The emitted energy has both polarizations r and l. A similar picture can be drawn for energy emitted due to energy in state l incident in the direction $-\theta$. The blackbody emits power equally in all directions and polarizations, so $I_l(+\theta) = I_l(-\theta)$. For the arbitrary incident state i, the following equality thus holds:

$$\sum_e (R_{ie}(-\theta) + T_{ie}(-\theta)) = \sum_e (R_{ie}(\theta) + T_{ie}(\theta)) \quad (3.4)$$

The two relations (3.3) and (3.4) have comparable forms when equation (3.3) is summed over emitted states e and equation (3.4) is summed over incident states i. In order for these two relations to be consistent, then

$$\sum_{e,i} (R_{ie}(\theta) - R_{ie}(-\theta)) = \sum_{e,i} (T_{ie}(\theta) - T_{ie}(-\theta)) = 0 \quad (3.5)$$

The requirement that the total reflectance due to all incident polarization states is reciprocal with respect to incident angle still holds:

$$\sum_{e,i} (R_{ie}(\theta) - R_{ie}(-\theta)) = 0 \quad (3.6)$$

To simplify the notation, the quantity $\Delta R_i = \sum_e (R_{ie}(\theta) - R_{ie}(-\theta))$ is defined. If there are only two possible polarizations (say $i=r$ and $i=l$), and $\Delta R_r(\theta) \neq 0$, equation (3.6) is still satisfied if

$$\Delta R_r(\theta) = -\Delta R_l(\theta) \quad (3.7)$$

This means reflection within one polarization can be nonreciprocal if reflection within at least one of the other polarizations is nonreciprocal. This equality will be demonstrated by a specific example in section 3.3.

The remainder of this chapter is organized in the following manner: In section 3.2 a theory is constructed for the reflectance and transmittance of an electromagnetic wave incident on a semi-infinite antiferromagnet in the presence of an applied field. This is done for arbitrary angles of incidence and arbitrary polarizations of the incident wave. In section 3.3, specific results are presented from numerical calculations of reflection from MnF_2 , a uniaxial antiferromagnet, which are used to show that nonreciprocal reflection can occur for certain polarizations of the incident wave. Linearly polarized incident waves are found to be reflected reciprocally with respect to the angle of incidence θ , but circularly polarized incident waves are not. It will be seen that the nonreciprocity of the circular reflectances will satisfy the equality (3.7).

3.2 Theory.

The geometry is defined in figure 3.1. The material occupies the $y > 0$ half space with the easy axis along z . An applied field also lies along the z axis. The plane of incidence makes an angle ϕ with the z axis and the incident wavevector \vec{k} makes an angle θ with respect to the outward normal from the surface. The incident wavevector is $\vec{k}^< = (k_x^<, k_y^<, k_z^<)$ and the incident magnetic field is \vec{H}^i . The reflected wavevector is $\vec{k}^r = (k_x^r, k_y^r, k_z^r)$ and the magnetic field is \vec{H}^r .

The dielectric properties of the material are determined by $\vec{\epsilon}$ and $\vec{\mu}$ as defined by equations (2.3) and (2.9). Plane wave solutions of the form $e^{i(\vec{k} \cdot \vec{x} - \omega t)}$ are assumed. In explicit matrix form, equation (2.11) then becomes

$$\begin{bmatrix} -\left(\frac{\epsilon_1}{\epsilon_2} k_y^2 + k_1^2\right) & (i\omega_0^2 \mu_2 \epsilon_1 + \frac{\epsilon_1}{\epsilon_2} k_x k_y) & k_x k_z \\ (i\omega_0^2 \epsilon_1 \mu_2 + \frac{\epsilon_1}{\epsilon_2} k_x k_y) & -\left(\frac{\epsilon_1}{\epsilon_2} k_x^2 + k_1^2\right) & k_z k_y \\ k_x k_z & k_z k_y & -(k_y^2 + k_2^2) \end{bmatrix} \vec{H} = 0 \quad (3.8)$$

k_1 and k_2 are given by

$$k_1^2 = k_z^2 - \omega_0^2 \epsilon_1 \mu_1 \quad (3.9)$$

$$k_2^2 = k_x^2 - \omega_0^2 \epsilon_2 \quad (3.10)$$

The dispersion relation of bulk polaritons propagating in the material is found by setting the determinant of this matrix to zero. In the special case $k_z=0$, the matrix simplifies considerably and an explicit expression for $k_y(\omega, k_x)$ is obtained:

$$k_y^2 = \omega^2 \epsilon_2 \left(\frac{\mu_1^2 - \mu_2^2}{\mu_1} \right) - k_x^2 \quad (3.11)$$

This equation describes the propagation of bulk polaritons in a direction perpendicular to the easy axis.¹⁶ Note that there is only one value of k_y for each k_x and ω .

When $k_z \neq 0$, the determinant of the matrix in (3.8) is a second order polynomial in $(k_y)^2$ so that k_y has two possible magnitudes at each frequency. A wave incident on the material can then transmit energy into two bulk polariton modes, each mode travelling in a different direction but with the same frequency as the incident wave.

Outside the material, where $y < 0$, the susceptibilities are uniform and describe a vacuum:

$$\epsilon_{ij} = \delta_{ij} \quad (3.12)$$

$$\mu_{ij} = \delta_{ij} \quad (3.13)$$

Substitution of (3.12) and (3.13) into the wave equation (3.7) gives the usual wave equation for H in free space. With plane wave solutions of the form $e^{i(\vec{k} \cdot \vec{x} - \omega t)}$, the free space dispersion relation is simply $\omega_0 = k^<$. Similarly, the reflected wavevector obeys $\omega_0 = k^r$.

The parallel components of the wavevector must match across the interface between the vacuum and the material. Thus

$$k_x = k_x^r = k_x^< \quad (3.14)$$

and

$$k_z = k_z^r = k_z^< \quad (3.15)$$

These two relations plus $\omega_0 = k^< = k^r$ allow us to write $\vec{k}^r = (k_x, -k_y^<, k_z)$.

The amplitudes of the transmitted and reflected waves can be found in terms of the incident wave by applying the usual electromagnetic boundary conditions. The existence of two waves in the material (corresponding to the two k_y 's) means that there are three undetermined fields in the problem: \vec{H}^r , and two transmitted fields \vec{H}_1 and \vec{H}_2 . In terms of components, this means there are nine amplitudes to be related through boundary conditions at $y=0$.

The usual Maxwell boundary conditions on tangential \vec{H} , normal \vec{B} , and tangential \vec{E} provide five independent equations. Four more equations are needed to determine the fields. It turns out that the boundary condition on continuity of normal \vec{D} across the interface can be expressed as a linear combination of the boundary conditions on tangential \vec{H} , so this does not yield any new relations. Instead, the remaining four equations are obtained from the equations of motion (3.8).

For arbitrary directions of propagation, the equations of motion (3.8) couple the components of the \vec{H} fields in the material. This means two of these three equations can be used to write H_x and H_y in terms of H_z . Since there are two \vec{H} fields in the material, there are four new relations involving the amplitude of the fields (two equations for each k_y).

Writing

$$\begin{bmatrix} H_x \\ H_y \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} H_z$$

(3.16)

A matrix equation for the coefficients A and B is obtained from (3.8):

$$\begin{bmatrix} -\left(\frac{\epsilon_1}{\epsilon_2} k_y^2 + k_1^2\right) & i\omega_0^2 \epsilon_1 \mu_2 + \frac{\epsilon_1}{\epsilon_2} k_x k_y \\ \frac{\epsilon_1}{\epsilon_2} k_x k_y - i\omega_0^2 \epsilon_1 \mu_2 & -\left(\frac{\epsilon_1}{\epsilon_2} k_x^2 + k_1^2\right) \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} k_x \\ k_y \end{bmatrix} k_z \quad (3.17)$$

(Note that any two of the three equations in 3.8 could have been chosen). One should emphasize that each k_y results in a unique set of coefficients A and B. These two sets are labelled by the subscripts "1" and "2": k_{y1} , A_1 , B_1 , H_{z1} , and k_{y2} , A_2 , B_2 , H_{z2} .

The five Maxwell boundary conditions are next written in the above notation.

Continuity of tangential \vec{H} across $y=0$ gives

$$H_x^i + H_x^r - A_1 H_{z1} - A_2 H_{z2} = 0 \quad (3.18)$$

$$H_z^i + H_z^r - H_{z1} - H_{z2} = 0 \quad (3.19)$$

Continuity of normal \vec{B} gives

$$H_y^i + H_y^r + i\mu_2(A_1 H_{z1} + A_2 H_{z2}) - \mu_1(B_1 H_{z1} + B_2 H_{z2}) = 0 \quad (3.20)$$

Continuity of tangential \vec{E} , written in terms of the \vec{H} fields via $c\nabla \times \vec{H} = \partial/\partial t \vec{D}$, gives the two relations

$$k_y (H_z^i - H_z^r) - k_z (H_y^i + H_y^r) - (k_{y1} H_{z1} + k_{y2} H_{z2} - k_{z1} B_1 H_{z1} - k_{z2} B_2 H_{z2})/\epsilon_1 = 0 \quad (3.21)$$

$$k_x(H_y^i + H_y^r) - k_y^<(H_x^i - H_x^r) - (k_x B_1 H_{z1} + k_x B_2 H_{z2} - k_{y1} A_1 H_{z1} - k_{y2} A_2 H_{z2})/\epsilon_2 = 0 \quad (3.22)$$

These five equations are sufficient to determine \vec{H}^r , \vec{H}_1 and \vec{H}_2 for a given \vec{H}^i , ω , ϕ and θ . For a numerical solution it is convenient to rewrite equations (3.18-3.22) in matrix form:

$$\begin{bmatrix} 1 & 0 & 0 & -A_1 & -A_2 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 1 & 0 & (i\mu_2 A_1 - \mu_1 B_1) & (i\mu_2 A_2 - \mu_1 B_2) \\ 0 & -k_z & -k_y^< & -(k_{y1} - k_z B_1)/\epsilon_1 & -(k_{y2} - k_z B_2)/\epsilon_1 \\ k_y^< & k_x & 0 & -(k_x B_1 - k_{y1} A_1)/\epsilon_2 & -(k_x B_2 - k_{y2} A_2)/\epsilon_2 \end{bmatrix} \begin{bmatrix} H_x^r \\ H_y^r \\ H_z^r \\ H_{z1} \\ H_{z2} \end{bmatrix} + \begin{bmatrix} H_x^i \\ H_z^i \\ H_y^i \\ k_y^< H_z^i - k_z H_y^i \\ k_x H_y^i - k_y^< H_x^i \end{bmatrix} = 0 \quad (3.23)$$

A numerical solution goes as follows: first those k_{y1} and k_{y2} that cause the determinant of (3.8) to vanish are found. Note that k_{y1} and k_{y2} must have positive real parts in order to represent waves travelling away from the surface. Next, the matrix equation (3.17) is solved for the vectors containing A and B. This is done once for k_{y1} and once for k_{y2} , resulting in values for A_1 , B_1 , A_2 , and B_2 . The matrix equation (3.23) is then solved for the vector containing \vec{H}^r , H_{z1} and H_{z2} . Finally, equation (3.16) is used to determine the x and y components of the two transmitted fields.

The last task is to calculate reflectances and transmittances. The reflectance is defined as the ratio of the reflected energy flux normal to the surface to the incident energy flux normal to the surface. Using the Poynting vector $\vec{S} = (\vec{E} \times \vec{H}^*) / 4\pi c$ this is simply

$$R = \frac{|\vec{H}^r|^2}{|\vec{H}^i|^2} \quad (3.24)$$

The transmittance is defined as the ratio of the transmitted energy flux normal to the surface to the incident energy flux normal to the surface. Since the material is gyrotropic, the energy flow is not in the direction of propagation. Also, since there are two waves in the material, there are also two transmittances. The transmittance then has the more complicated form:

$$T_n = |\vec{H}^i|^2 \frac{c}{\epsilon_1 k_y} \text{Re} \left[|\vec{H}_n|^2 k_{yn} - (\vec{k}_n \cdot \vec{H}_n^*) H_{yn} \right] \quad (3.25)$$

The subscript "n" can have the value 1 or 2 and identifies the corresponding transmitted field and wavevector according to the previous convention.

It is convenient for the following discussion to represent the incident and reflected H fields in terms of components parallel and perpendicular to the plane of incidence. The incident field's parallel component is defined as H_p^i , and its perpendicular component is defined as H_s^i . The transformation from incident parallel and perpendicular components to cartesian components is given by

$$H_x^i = H_p^i \cos\theta \sin\phi + H_s^i \cos\phi \quad (3.26)$$

$$H_y^i = -H_p^i \sin\theta \quad (3.27)$$

and

$$H_z^i = H_p^i \cos \theta \cos \phi - H_s^i \sin \phi \quad (3.28)$$

Since the theory calculates the reflected fields in terms of their cartesian components, the reverse transformation is required. Again referring to the geometry of figure 3.1, the parallel (H_p^r) and perpendicular (H_s^r) components of the reflected fields are given by

$$H_s^r = H_x^r \cos \phi - H_z^r \sin \phi \quad (3.29)$$

$$H_p^r = -H_y^r / \sin \theta \quad (3.30)$$

3.3 Results.

The preceding theory is now applied to the specific example of reflection from the antiferromagnet MnF_2 . This is a uniaxial antiferromagnet with the parameters $H_{\text{ex}} = 550$ kG, $H_a = 7.87$ kG, and $M_s = .6$ kG. Damping is ignored. The following results are presented in the unitless variables ω/Ω and ck/Ω where Ω is the antiferromagnetic resonance frequency defined in Chapter 2.

The primary goal is to establish conditions for nonreciprocal reflection with respect to the incident angle θ ; i.e., conditions such that $R(\theta) \neq R(-\theta)$. As discussed in the introduction, the two propagation directions $+k$ and $-k$ (corresponding to $+\theta$ and $-\theta$) are equivalent when there is no applied field. It is only when an applied field is present that $+k$ and $-k$ propagation directions can lead to nonreciprocal properties. A static external field of .3kG is thus applied in the z direction.

In this applied field, electromagnetic waves can propagate in MnF_2 if they have frequencies below $\omega/\Omega = .998$, between $\omega/\Omega = 1.002$ and $\omega/\Omega = 1.009$, or above $\omega/\Omega = 1.012$. These frequency regions are often referred to as the bulk polariton bands and are defined as

the regions where the normal component of the wavevector in the material (equation 2.5) is real.

In figure 3.3, the reflectance R and the two transmittances T_1 and T_2 are plotted as functions of incident angle for a linearly polarized incident wave. The incident H field lies in the plane defined by $H_s^i = H_p^i$ and the frequency is $\omega/\Omega = .990$, well within the lower bulk polariton band. The plane of incidence makes an angle $\phi = 45^\circ$ from the z axis. Note that most of the transmitted energy has gone into the T_2 bulk mode although at angles near $\theta = 90^\circ$, the T_1 bulk mode is also excited.

Most importantly, however, $R(\theta) = R(-\theta)$ so the reflectance is reciprocal. This reciprocity occurs for all angles ϕ and all orientations of the plane of polarization of the incident wave. The only difference for different orientations of the plane of polarization of the incident wave is in the relative amounts of energy transmitted into the T_1 and T_2 bulk modes. T_1 has its maximum for the above case where $H_s^i = H_p^i$ and T_2 is maximum for the case $H_s^i = -H_p^i$.

A very different situation occurs when the incident wave is circularly polarized. In figure 3.4, $\Delta R = R(\theta) - R(-\theta)$ is shown for a left circularly polarized incident wave where $H_p^i = iH_s^i$. The frequency is $\omega/\Omega = .990$ and the applied field is .3kG. The reflectance is clearly nonreciprocal with respect to θ . The magnitude of ΔR increases with θ until it reaches a maximum near $\phi_0 = 75^\circ$. At $\theta = 90^\circ$, the incident wave travels parallel to the surface and R vanishes. Consequently, beyond ϕ_0 , ΔR must also tend to zero.

This nonreciprocal behavior can be understood through the following considerations. First, recall that magnetic waves in the material drive the magnetizations of the sublattices through Bloch's equations of motion. This results in a precession of the magnetizations about the z axis and gives rise to magnetic fields which rotate in the xy plane. One might then expect this rotation to affect the polarization of the bulk modes.

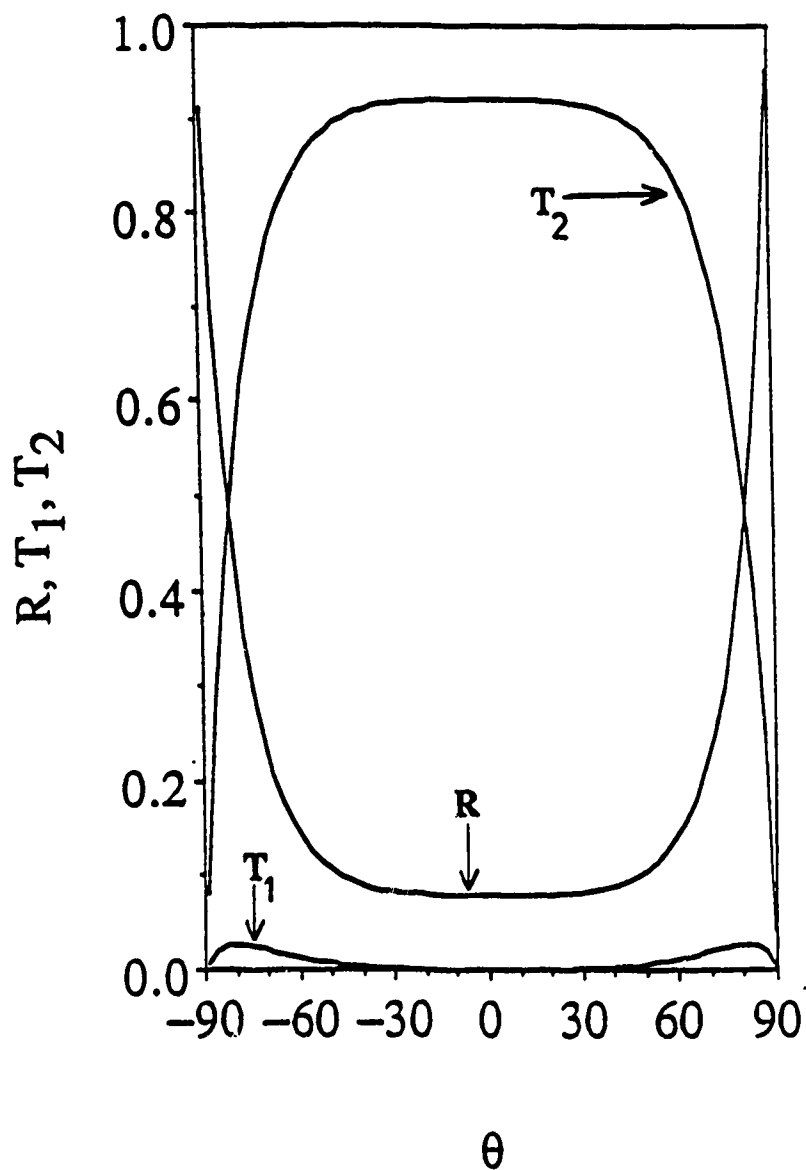


Figure 3.3. Reflectance and transmittance for a linearly polarized wave ($H_s^i = H_p^i$) incident at $\phi = 45^\circ$ as functions of incident angle θ . The frequency is $\omega/\Omega = 0.990$. The two transmittances, T_1 and T_2 , correspond to the two transmitted waves. The reflectance for $+\theta$ is identical to that of $-\theta$.

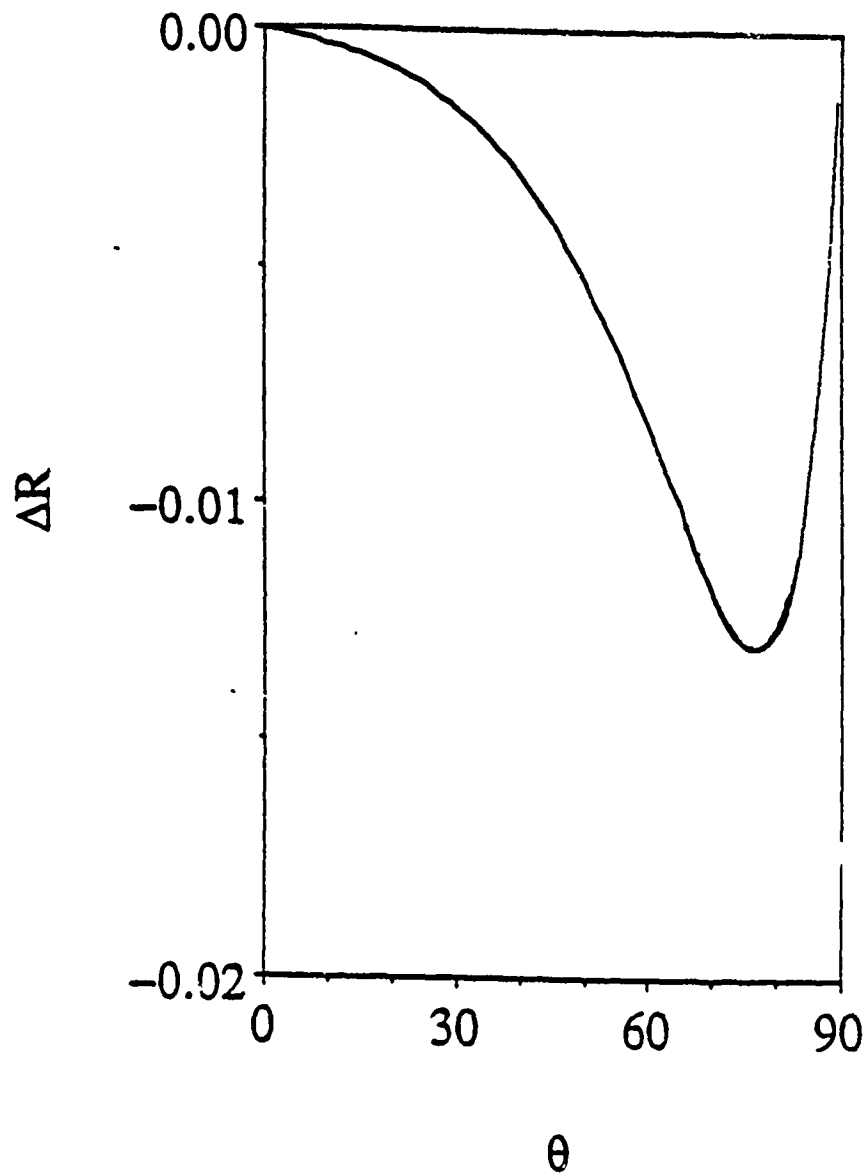


Figure 3.4. The change in reflectance, $\Delta R(\theta) = R(\theta) - R(-\theta)$, for a left circularly polarized wave ($H_s^1 = iH_p^1$) incident at $\theta = 45^\circ$ as a function of incident angle θ . The frequency is $\omega/\Omega = 0.990$. Here $\Delta R(\theta)$ is negative implying that the reflectance for $-\theta$ is greater than that of $+\theta$.

This in turn may mean the bulk modes are most easily excited by fields with certain polarizations, thus leading to a dependence of the transmittance on the polarization of the driving, or incident, wave. This is pursued by examining the polarization of the bulk polariton modes through the equations of motion (3.8).

Consider propagation in the yz plane where $k_x=0$. As long as $H_0 \neq 0$, the three components of the bulk fields H_x , H_y , and H_z are coupled together through μ_2 . The components are related as follows:

$$H_x = i \frac{1}{k_z} \rho(k_z) \eta(k_z) H_z \quad (3.32)$$

and

$$H_y = \frac{1}{k_z} \eta(k_z) H_z \quad (3.33)$$

The quantities $\rho(k_z)$ and $\eta(k_z)$ are given by

$$\rho(k_z) = \frac{\omega_o^2 \epsilon_1 \mu_2}{(\epsilon_1 / \epsilon_2) k_y^2 + k_1^2} \quad (3.34)$$

and

$$\eta(k_z) = k_y \frac{\omega_o^2 \epsilon_2}{k_y} \quad (3.35)$$

The fields of a bulk mode in the vector form are

$$\vec{H}(k_z) = (i\rho(k_z)\eta(k_z)/k_z, \eta(k_z)/k_z, 1) H_z(k_z) \quad (3.36)$$

This represents an elliptically polarized wave whose magnetic field rotates in both the xy and xz planes. Now let k_z go to $-k_z$ and observe the resulting polarization.

First, note that the normal component of the wavevector, k_y , is determined by setting the determinant of the equations of motion equal to zero, as before. This results in a polynomial which is a quadratic in k_z^2 . Thus $k_y(k_z) = k_y(-k_z)$. The coefficients $\rho(k_z)$ and $\eta(k_z)$ depend only on k_z^2 and k_y , so $\rho(k_z) = \rho(-k_z)$ and $\eta(k_z) = \eta(-k_z)$. Finally, from equation (3.28) it is seen that H_z is unchanged as θ goes to $-\theta$ (which is equivalent to letting k_z go to $-k_z$). The resulting bulk \vec{H} field is

$$\vec{H}(-k_z) = (-i\rho(k_z)\eta(k_z)/k_z, -\eta(k_z)/k_z, 1) H_z(k_z) \quad (3.37)$$

The wave is still elliptically polarized but the direction of rotation in the xz plane is reversed from $\vec{H}(k_z)$. The direction of rotation in the xy plane, however, is the same as that of $\vec{H}(k_z)$.

To see how this might lead to nonreciprocal transmission and reflection, suppose the incident wave is travelling in the yz plane with $+k_z$ and is circularly polarized with x and y components $H_x^i = 1$ and $H_y^i = i$. An identical circularly polarized wave travelling in the yz plane but with $-k_z$ would need to have $H_x^i = 1$ and $H_y^i = -i$ in order to preserve the sense of rotation of the field about the wave's direction of propagation.

An incident wave with a given sense of rotation should find it easier to excite waves in the material that have the same sense of rotation. A measure of the incident wave's ability to excite bulk modes in the material is the magnitude of the product $(\vec{H}^i)^* \cdot \vec{H}$.

Using the above circularly polarized incident wave and arbitrarily setting the z component to unity for both the $+k_z$ and $-k_z$ directions, one obtains for the $+k_z$ direction:

$$|\vec{H}^i(k_z) \cdot \vec{H}(k_z)| = H_z(k_z) \sqrt{1 + (1 + \rho(k_z))^2 \frac{\eta^2(k_z)}{k_z^2}} \quad (3.38)$$

For the $-k_z$ direction, the corresponding expression is:

$$|\vec{H}^i(-k_z) \cdot \vec{H}(-k_z)| = H_z(k_z) \sqrt{1 + (1 - \rho(k_z))^2 \frac{\eta^2(k_z)}{k_z^2}} \quad (3.39)$$

Suppose ρ is negative for one of the transmitted waves. Equation (3.38) is then less than (3.39) and the transmittance into this wave should be greater for incidence with $-k_z$ than with $+k_z$. If the other bulk polariton also has negative ρ , or simply has smaller magnitudes for ρ and η , then the reflectance will be greatest for incidence with $+k_z$.

In the example of figure 3.4, an examination of the numerically calculated A and B coefficients show that the bulk polariton modes corresponding to each value of k_y have opposite senses of rotation and different magnitudes in the xy plane. The bulk polariton with the largest x and y components in this case is right circularly polarized for $-k_z$. Thus the transmittance is greater for $+k_z$ than $-k_z$ and the reflectance is less for $+\theta$ than for $-\theta$.

On the other hand, a right circularly polarized incident wave should reverse the situation and transmit more energy into the bulk modes at $-k_z$ than at $+k_z$. This is seen in figure 3.5 where $\Delta R = R(\theta) - R(-\theta)$ is shown for a right circularly polarized incident wave at $\omega/\Omega = .990$ and $\phi = 45^\circ$. Here ΔR is positive, indicating a greater transmittance at $-\theta$ than at $+\theta$.

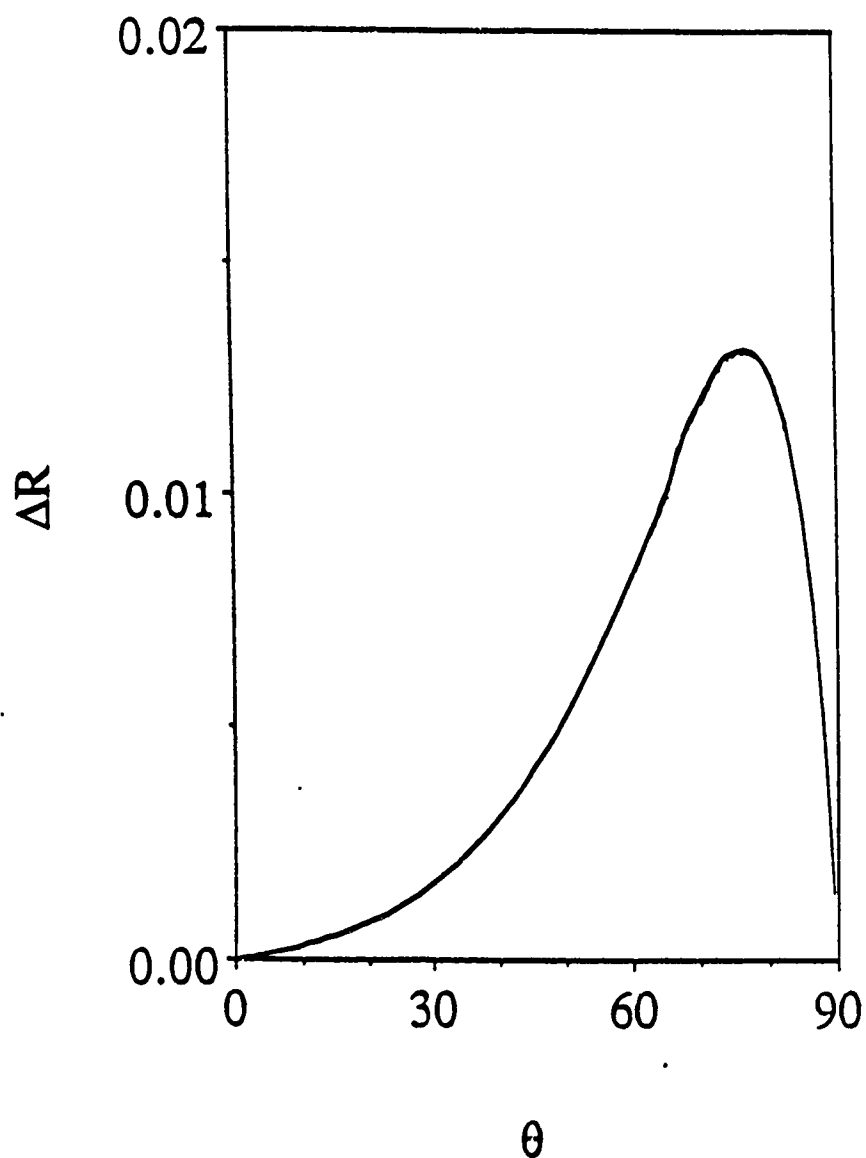


Figure 3.5. The change in reflectance, $\Delta R(\theta) = R(\theta) - R(-\theta)$, for a right circularly polarized wave ($H_s^1 = iH_p^1$) incident at $\phi = 45^\circ$ as a function of incident angle θ . The frequency is again $\omega/\Omega = .990$. Here $\Delta R(\theta)$ is positive implying that the reflectance for $-\theta$ is less than that of $+\theta$.

From figures 3.5 and 3.6 it is clear that the $\Delta R(\theta)$ for left circular incident light is exactly equal and opposite the $\Delta R(\theta)$ for right circularly polarized incident light. This is precisely the prediction of the thermodynamic argument of section 3.1. Since linearly polarized light consists of equal amounts of right and left circularly polarized light, the net change in reflectance, $\Delta R_T(\theta) + \Delta R_L(\theta)$ must vanish by the same argument. This explains why the linearly polarized incident wave of figure 3.3 has the same reflectance for $+\theta$ and $-\theta$.

If the nonreciprocity of a circularly polarized incident wave depends on the magnitude of its x and y field components, then ΔR should depend on ϕ . The magnitudes of these components are largest for small ϕ and large θ . In figure 3.6 ΔR for $\phi=10^\circ$, 45° , and 80° is plotted for a right circularly polarized incident wave of frequency $\omega/\Omega=.995$. It is clear that ΔR increases as ϕ approaches 0. Also, in all three cases ΔR is largest for large θ .

Note that ΔR is much larger for all three of the cases of figure 3.6 due to the higher frequency of the incident wave. The frequency $\omega/\Omega=.995$ is much nearer the antiferromagnetic resonance frequency so μ_2 is much larger than at $\omega/\Omega=.990$. This means an increased coupling between H_x and H_y for $\omega/\Omega=.995$ and a consequent increase in the nonreciprocal behavior of the reflectance. In regions where μ_1 and μ_2 are large, the rotation direction of the bulk modes as well as their magnitudes in the xy planes can change drastically. This can allow a wave incident at one $+\theta$ to drive one of the bulk modes strongly while at another still positive θ , the other bulk mode is driven strongly.

The fact that the two transmitted waves have opposing directions of rotation, and that their magnitudes depend on frequency, allows the change in reflectance, ΔR , to change sign at different frequencies and angles of incidence. In figure 3.7 ΔR is plotted for right circularly polarized light at $\omega/\Omega=1.006$ and $\omega/\Omega=1.015$ with $\phi=45^\circ$. In both cases, μ_1 and μ_2 are relatively large so that the nonreciprocal reflection is also well pronounced. Here note the changing sign of ΔR . At small θ , ΔR is small and positive and at large θ , ΔR is large and negative.

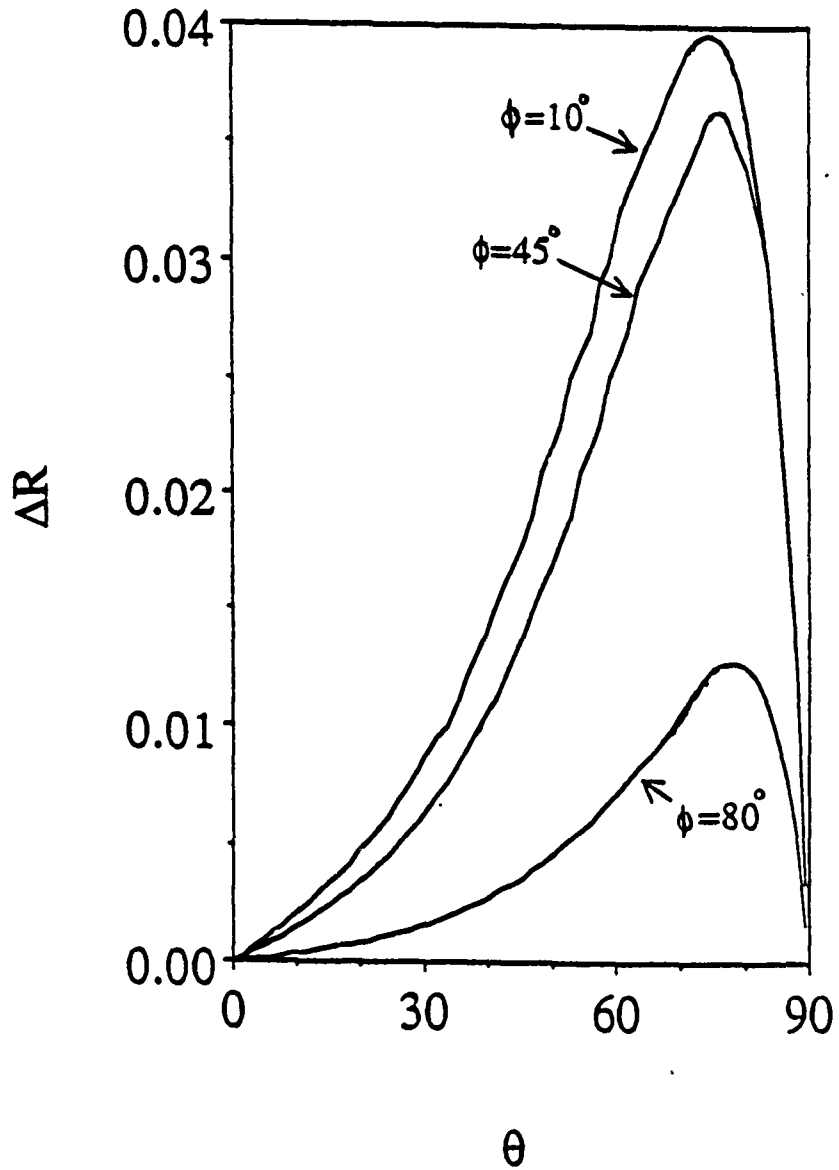


Figure 3.6. The change in reflectance, $\Delta R(\theta) = R(\theta) - R(-\theta)$, for a right circularly polarized wave ($H_s^1 = iH_p^1$) incident at $\phi = 10^\circ$, 45° , and 80° as functions of incident angle θ . The frequency is $\omega\Omega = 995$. $\Delta R(\theta)$ is largest for incident waves travelling along the direction of the magnetic field. Note that $\Delta R(\theta)$ is much larger at this frequency than at $\omega\Omega = 990$.

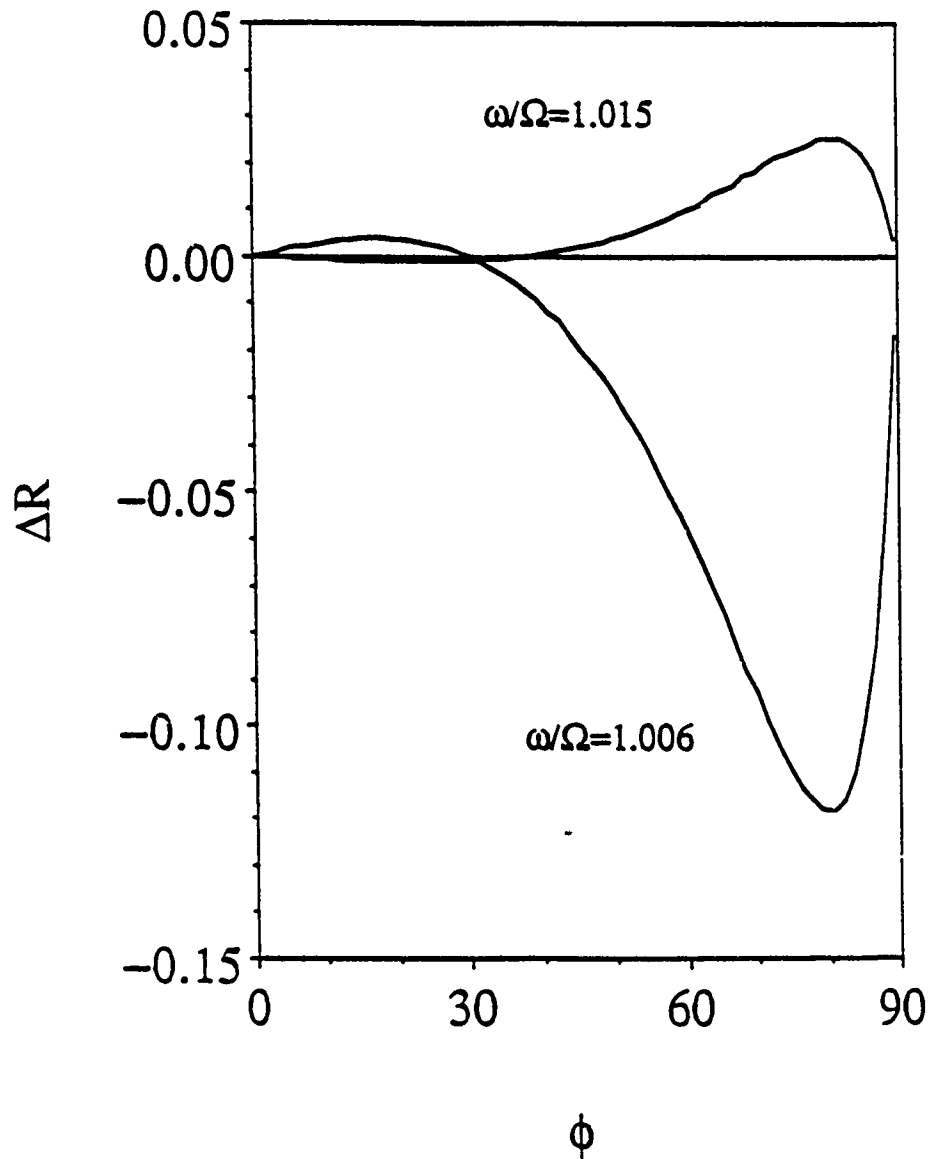


Figure 3.7. The change in reflectance, $\Delta R(\theta) = R(\theta) - R(-\theta)$, for a right circularly polarized wave ($H_s^i = iH_p^i$) incident at $\phi = 45^\circ$ as a function of incident angle θ . The frequencies are $\omega/\Omega = 1.006$ and $\omega/\Omega = 1.015$. These two frequencies lie within the middle and upper bulk polariton bands, respectively. Note that $\Delta R(\theta)$ has both positive and negative values for each case, corresponding to the complicated behavior of the bulk modes near the antiferromagnetic resonance frequencies.

The strong nonreciprocity in R coincides with a strong nonreciprocity in the transmittances. This is illustrated in figure 3.8 where R, T_1 and T_2 are plotted as functions of θ for the $\omega/\Omega=1.006$ case of figure 3.7. The relative magnitudes of the two transmitted waves are reversed as θ goes to $-\theta$, indicating the opposite rotation directions of the two waves.

The large difference between the magnitudes of the two transmittances of figure 3.8 indicates that the largest ΔR should be found by examining separately the left and right circularly polarized components of the reflected wave. If the reflected wave $\vec{H}^r = (H_s^r, H_p^r)$ is written in terms of right and left handed circularly polarized unit vectors \hat{r} and \hat{l} , then a simple transformation from the (s,p) basis to the (r,l) basis yields

$$\vec{H}^r = a\hat{r} + b\hat{l} \quad (3.40)$$

The squared magnitudes of the complex coefficients a and b are the reflectances of the right and left circularly polarized components. These are given by

$$R_r = a^*a / (a^*a + b^*b) = [|H_s^r|^2 + |H_p^r|^2 + 2|H_s^r||H_p^r|\sin(\alpha_p - \alpha_s)] / 2|\vec{H}^r|^2 \quad (3.41)$$

and

$$R_l = b^*b / (a^*a + b^*b) = [|H_s^r|^2 + |H_p^r|^2 - 2|H_s^r||H_p^r|\sin(\alpha_p - \alpha_s)] / 2|\vec{H}^r|^2 \quad (3.42)$$

The phases α_p and α_s are defined by

$$\tan \alpha_p = \text{Im}(H_p^r) / \text{Re}(H_p^r) \quad (3.43)$$

and

$$\tan \alpha_s = \text{Im}(H_s^r) / \text{Re}(H_s^r) \quad (3.44)$$

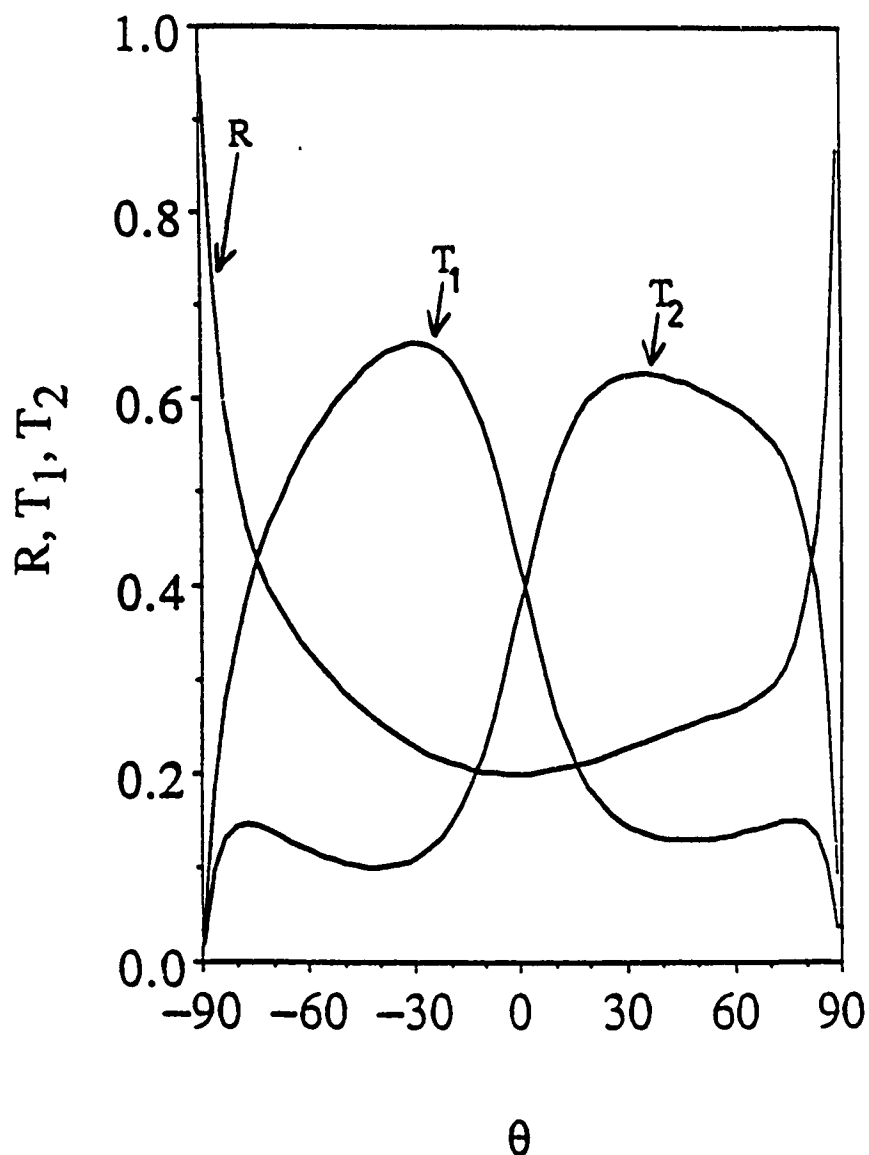


Figure 3.8. Reflectance and transmittance for a right circularly polarized wave ($H_s^i = iH_p^i$) incident at $\phi = 45^\circ$ as functions of incident angle θ . The frequency is $\omega\Omega = 1.006$. The two transmittances, T_1 and T_2 , correspond to the two transmitted waves. T_1 is much larger than T_2 for $-\theta$ while T_2 is larger than T_1 for $+\theta$. The net transmittance, and hence reflectance, is still nonreciprocal, however.

A comparison of $\Delta R_T(\theta) = R_T(\theta) - R_T(-\theta)$ and $\Delta R_I(\theta) = R_I(\theta) - R_I(-\theta)$ shows that the nonreciprocity is due largely to scattering within one polarization. At $\omega/\Omega = 1.006$, for example, $\Delta R_I(\theta)$ is nearly zero for all θ and $\Delta R(\theta) \approx \Delta R_T(\theta)$.

The primary purpose of this chapter has been to show that nonreciprocal reflection can exist without absorption by the material. In practical considerations, absorption is of course always present to some degree, and it is of interest to examine the effects of damping on the nonreciprocity of the reflectance. When Bloch damping is included, it is found that ΔR is increased a small amount. The most noticeable effect, however, is the shifting of the angle where $\Delta R = 0$ for frequencies where ΔR changes sign. For a damping value of $1/\tau\Omega = .0002$, $\omega/\Omega = 1.006$ and an applied field of .3kG, this shift is 3° .

CHAPTER 4

POLARITONS WITH DAMPING

The goal of this chapter is to discuss the effects of damping on polariton modes in antiferromagnets. There are several reasons for this. Damping often plays a critical role in the interaction of external probes with polaritons. For example, in the reflection experiment mentioned earlier, damping is necessary for a coupling of the external electromagnetic radiation to the surface polariton modes. Also, as discussed in chapter 2, not only does damping have a significant impact on the surface polariton modes, but the inclusion of damping also leads to the existence of new "leaky" surface modes.

This chapter begins with the effects of damping on the semi-infinite antiferromagnetic polariton curves and discusses the properties of the magnetic Brewster mode. The effects on the leaky modes from using an alternative damping mechanism is also considered.

4.1 Surface polaritons with Bloch-Bloembergen damping.

A more realistic description of surface polaritons is obtained by including a damping mechanism in the equations of motion. In general, damping affects the degree of localization of the wave to the surface and causes the wave to lose energy as it propagates. Although damping and its effects on surface polaritons has been studied for plasmon polaritons and magnetoelastic polaritons on ferromagnets, antiferromagnetic surface polariton studies have only considered the idealized case of zero damping. In this section damping is included in

the description and the resulting effects on the surface polariton's dispersion curve, path length and penetration depth are examined.

In addition to modifying the properties of the "true" antiferromagnetic surface polaritons, the inclusion of damping shows the existence of new surface resonances, or "leaky" surface modes. These are very different in character from the modes found without damping.²² The "true" surface modes and the leaky modes are investigated by numerically examining the surface polariton dispersion relation (2.15) for the material MnF_2 with the parameters $H_e=550\text{kG}$, $H_a=7.87\text{kG}$, $M=.6\text{kG}$ and $\epsilon_2=5.5$.

As discussed in chapter 2, one obtains the dispersion relation by assuming plane wave solutions of the form $\exp(i(k_x x - \omega t))\exp(-\alpha y)$ inside the material and $\exp(i(k_x x - \omega t))\exp(\gamma y)$ outside the material (the geometry is the same as that used in the previous chapters with the material in the $y>0$ half space). The fields inside and outside the material are then matched according to Maxwell's boundary conditions on tangential H and normal B. The resulting implicit expression was derived by Camley and Mills:⁶

$$\gamma + (\mu_1 \alpha + \mu_2 k_x) / (\mu_1^2 - \mu_2^2) = 0 \quad (4.1)$$

k_x is again the wavevector parallel to the surface and decay in the y direction is governed by α and γ (defined in (2.13) and (2.14)) which are given by:

$$\alpha = \sqrt{k_x^2 - \omega^2 \epsilon_2 \left(\frac{\mu_1^2 - \mu_2^2}{\mu_1^2} \right)} \quad (4.2a)$$

and

$$\gamma = \sqrt{k_x^2 - \omega^2} \quad (4.2b)$$

When there is no damping in the material, equation (4.1) can be satisfied by appropriate choices of real frequencies and real wavevectors. In this case both α and γ are positive and real. These modes represent excitations that are bound to the surface, with infinite lifetimes. Also, when $H_0=0$ the modes are reciprocal in k . This means the solutions obey $\omega(k)=\omega(-k)$. In the presence of an applied field the modes are not reciprocal and the solutions obey $\omega(k)\neq\omega(-k)$.

Bulk modes, on the other hand, exist in frequency ranges where α is pure imaginary. One sees from (4.2) that these ranges exist for those ω, k_x such that $\alpha^2 < 0$. In an infinite geometry, all values of α are allowed and so the number of bulk modes is infinite in each bulk band. Also, the bulk modes are reciprocal in k_x both with and without an applied field.

When damping is present in the magnetic system (through τ in the susceptibilities of (2.3) and (2.4)), the dispersion equation (4.1) no longer possesses pure real wavevector and frequency solutions. The dispersion relation can be satisfied, however, for real frequencies and complex wavevectors. With complex α , γ and k_x , these solutions represent dissipative waves that have finite path lengths.

To illustrate the properties of these dissipative waves, equation (4.1) was solved numerically for MnF_2 both with and without damping. The quantities plotted are unitless with reduced frequencies of ω/Ω and reduced wavevector $k_x c/\Omega$ (the component parallel to the surface).

In figure 4.1 the dispersion relations for bulk and surface polaritons are reproduced for the case of $H_0=0$. The shaded areas represent the bulk modes and the dashed lines between the two bulk bands are surface modes for the case of zero damping. Note that these surface modes stop abruptly at the top of the lower bulk band where $k_x=\omega/c$.

As the wavelength decreases, ω/k goes to zero and the surface polaritons asymptotically approach the zero field magnetostatic surface wave frequency given by

$$\omega_s = \gamma [H_a(2H_e + H_a + 4\pi M)]^{1/2}.$$

In figure 4.1 the surface polariton solutions to (4.1) when $1/\omega\tau = .0001$ are also included. These solutions are plotted as functions of reduced frequency and the real part of the reduced parallel wavevector. These dissipative waves lie very near the $1/\Omega\tau = 0$ surface polaritons in the frequency region above the lower bulk band and below the magnetostatic frequency. Outside this region, new solutions appear and are represented by solid lines.

Above ω_s the dissipative waves exhibit a curious "backbending" property where the group velocity changes sign and the modes curve inward in k_x toward the upper bulk band. This back-bending effect is reminiscent of a similar behavior found for surface plasmon-polaritons where the Fano ("true" surface modes) bend back with increasing frequency into what are sometimes called evanescent modes.²² The evanescent modes are tightly bound to the surface and the real and imaginary parts of the wavevector component normal to the surface have roughly the same magnitude. It is interesting to note that since the point of bend-back occurs near ω_s , the corresponding k_x can be used to measure the damping parameter. As the frequency increases above ω_s , the real parts of γ and α become small so that the evanescent modes are less tightly bound to the surface. Near the lower limit of the upper bulk band, the real parts of γ and α tend to zero and the evanescent modes become ill-defined.

With damping present, the polariton mode continues into the lower bulk band below the lower bulk band frequency limit, as seen in figure 4.1. This is a region forbidden to true surface polaritons. In this frequency region modes can exist, with damping present, at an ω and k_x for which a wave incident on the material from vacuum would be completely transmitted. We thus identify the $1/\Omega\tau = .0002$ mode as the magnetic analogue of the Brewster mode found in dielectric materials.

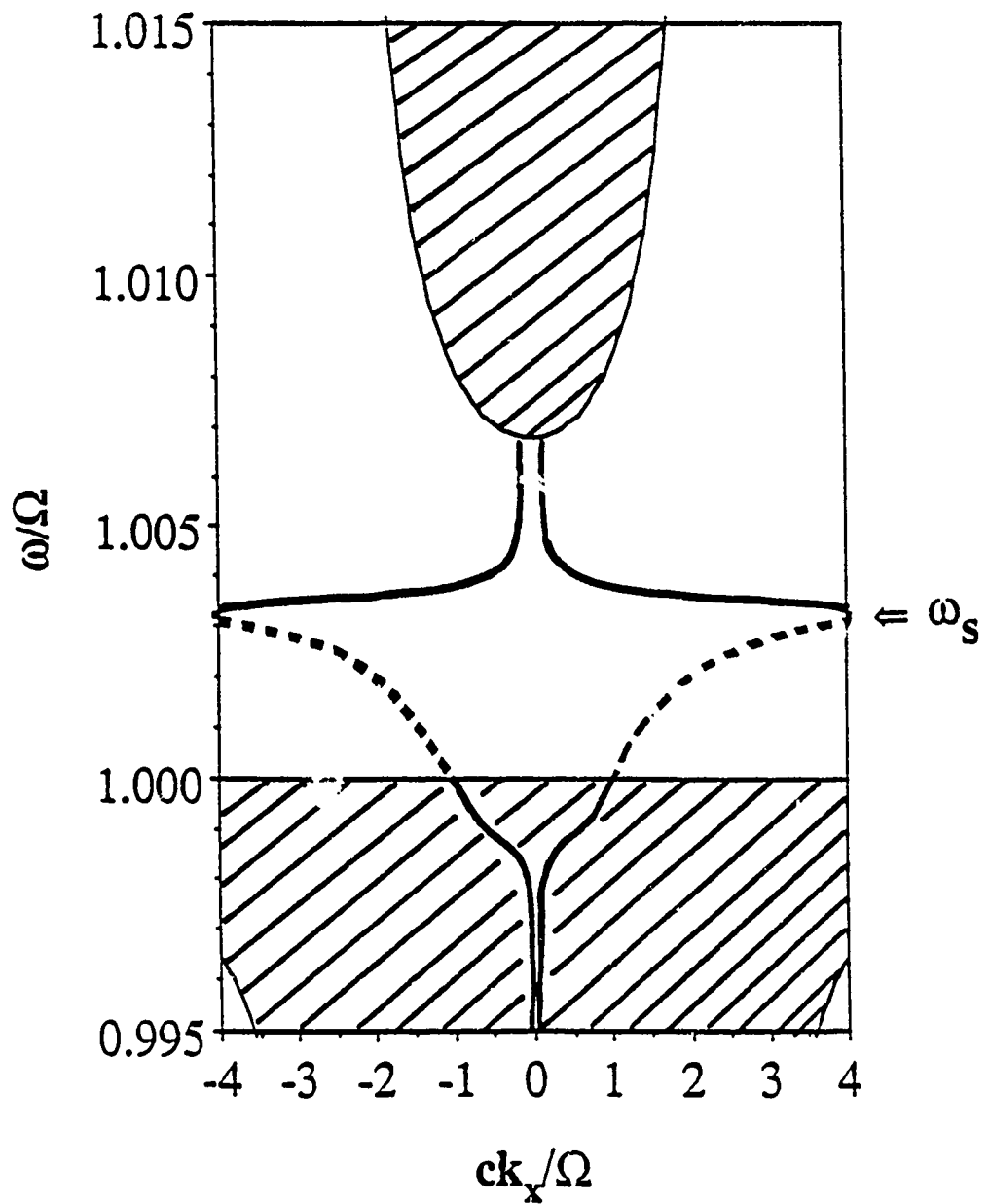


Figure 4.1. Dispersion curves for antiferromagnetic polaritons, in MnF_2 , with no applied field. The shaded areas are bulk bands and the dashed lines are the surface modes when there is no damping present. The solid lines show how damping modifies the surface mode frequencies. These solid lines are the real parts of the complex solutions to the dispersion relation (4.1), with damping $1/\Omega\tau = 0.0002$.

Typically, one thinks of the Brewster angle as the angle of incidence where there is no reflected wave from a surface illuminated by a wave polarized with its electric field in the plane of incidence. A similar angle occurs for a wave incident on a magnetic material and polarized with its magnetic field in the plane of incidence. This angle (or rather the corresponding component of the incident wavevector, k_x) can be found by setting to zero the appropriate Fresnel relations for the amplitude of the magnetic field of the reflected wave (see appendix A for the Fresnel relations). Doing so, one arrives at the magnetic Brewster condition:

$$(k_x)_{Br} = \omega_0 [\mu_1 \epsilon_2 (\mu_1^2 - \mu_2^2) - (\mu_1^2 - \mu_2^2)^2] / [\mu_1^2 - (\mu_1^2 - \mu_2^2)^2]^{1/2} \quad (4.3)$$

In figure 4.2 the ω and k_x that satisfy equation (4.3) are plotted along with the solutions to the dispersion relation (4.1) for $1/\Omega\tau = .0001$. The real part of the complex wavevector solutions are plotted against the frequency both with an applied field and without. When there is no applied field there is a very close correspondence between the two curves. When $H_0 = .3\text{kG}$, however, the two curves begin to differ for frequencies well within the lower bulk band.

Since $\text{Re}(\alpha)$ determines how tightly the surface wave is bound to the surface and $\text{Im}(k_x)$ governs the attenuation of the wave as it propagates parallel to the surface, it is interesting to plot these quantities as functions of frequency. In figure 4.3 $|\text{Im}(ck_x/\Omega)|$ is shown, as determined from the dispersion (4.1), and the corresponding $\text{Re}(c\alpha/\Omega)$ for zero applied field and $1/\Omega\tau = .0002$. The solid line is $|\text{Im}(k_x)|$ and the dashed line is $\text{Re}(c\alpha/\Omega)$. $\text{Im}(k_x)$ becomes large only in the bulk band and near ω_g . In the surface mode

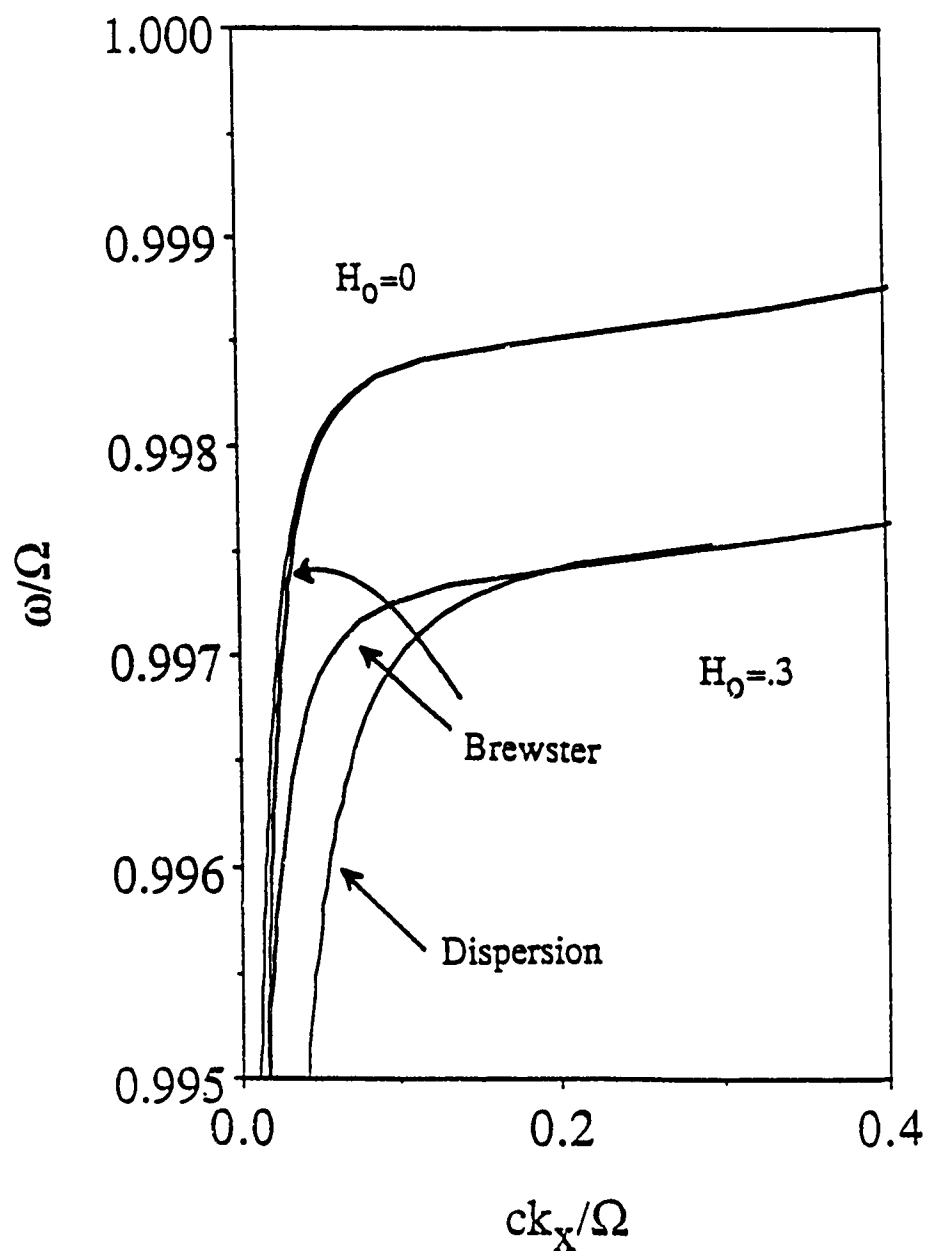


Figure 4.2. Real part of the complex dispersion solutions compared to the Brewster angle (in terms of k_x rather than θ_0). The upper curves are for $H_0=0$ and the lower curves are for the $-k_x$ branch of the surface polariton dispersion with $H_0=.3\text{kG}$. In the $H_0=0$ case, the curves are identical at higher frequencies.

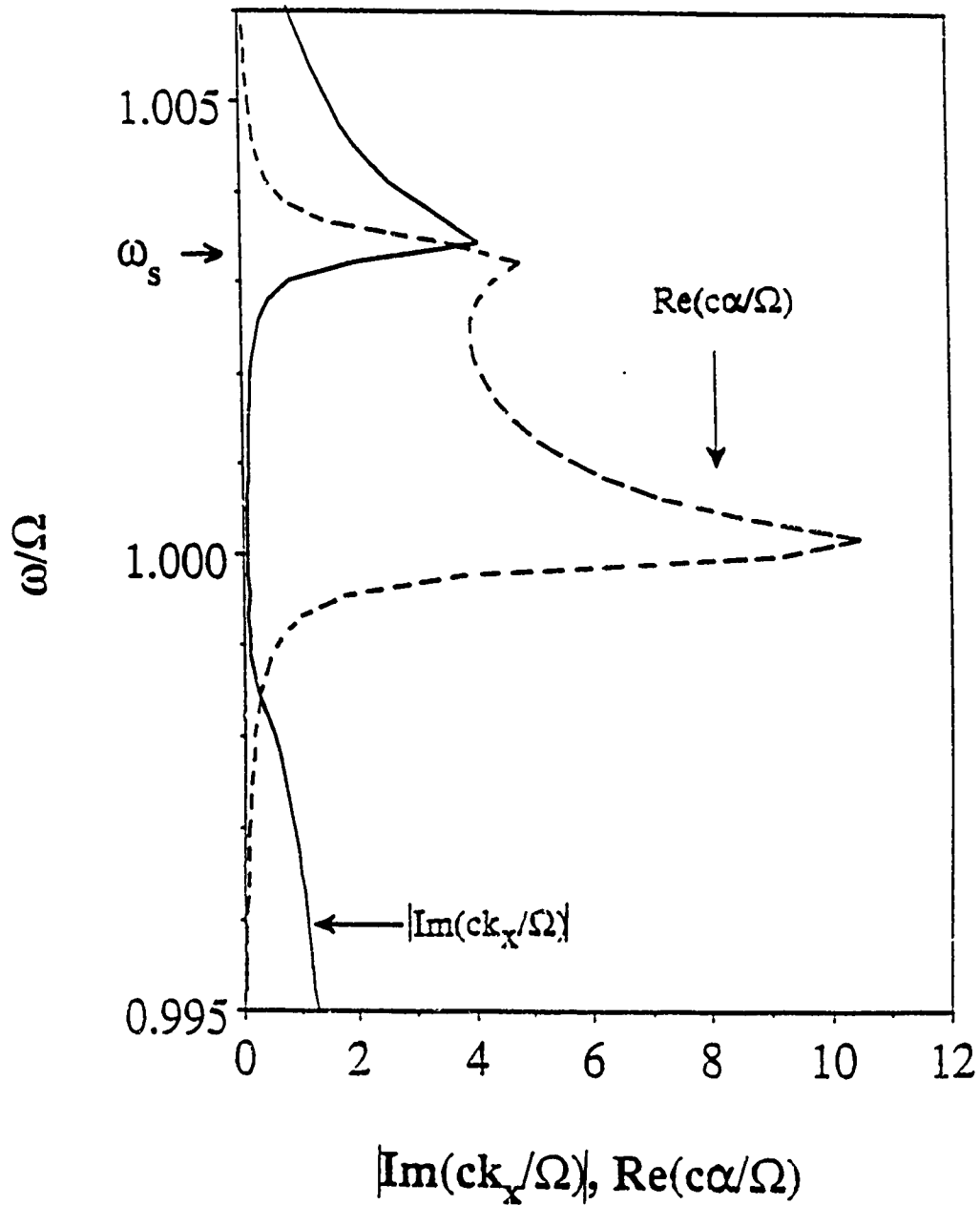


Figure 4.3. Decay parameter $|\text{Im}(ck_x/\Omega)|$ and localization parameter $\text{Re}(c\alpha/\Omega)$ for the $+k_x$ branch of the $1/\Omega\tau=0.0002$ dispersion curve of figure 4.1. Only the magnitudes are shown. The solid line is $|\text{Im}(k_x)|$ and the dashed line is $\text{Re}(c\alpha/\Omega)$. Note the strong localization to the surface just above $\omega/\Omega=1$. The localization is greatest in the surface mode region near the lower bulk band frequency limit.

region, $\text{Im}(k_x)$ is small so the wave has a long path length. $\text{Re}(\alpha)$ is very large for frequencies below ω_s and above the bulk band, thus indicating a strong localization to the surface. The localization is strongest near the bulk band limit, $\omega/\Omega=1$. In the bulk band, however, $\text{Re}(\alpha)$ is small and so the mode in this region is weakly bound to the surface.

Damping allows the surface mode to lose energy into the material. In figure 4.4, the $+k_x$ branch of the dispersion curve plus the decay parameters $\text{Im}(k_x)$ and $\text{Re}(\alpha)$ are plotted for $1/\omega\tau=.0008$. The solid line is $|\text{Im}(k_x)|$ and the dashed line is $\text{Re}(c\alpha/\Omega)$. H_0 is still zero. With greater damping the surface mode is not as tightly bound to the surface as before and penetrates further into the material. This increases the rate of energy loss into the material and thus the surface mode has a shorter path length. In this way damping allows the modes to "leak" energy into the interior of the material.

In an applied field, the surface modes outside the bulk bands are highly nonreciprocal. The resonances in the bulk bands are also nonreciprocal in applied fields, although the nonreciprocity disappears at frequencies below the lower bulk band limit. This is seen in figure 4.5 where the real part of k_x is plotted against frequency for $H_0=.3\text{kG}$. The dashed lines are the $1/\Omega\tau=0$ modes and the solid lines are the dissipative waves for $1/\Omega\tau=.0002$. The $1/\Omega\tau=0$ modes coincide with the $1/\Omega\tau=.0002$ modes except near the magnetostatic limit frequency. Again there is a close correspondence between the $1/\Omega\tau=0$ modes and the modes with damping in the surface mode region between the bulk band and the magnetostatic limit.

With the applied field, there are two regions where the real parts of γ and α become very small and the leaky modes become ill-defined. One region is above $\omega/\Omega=1.0075$, near the upper limit of the middle bulk band. The leaky modes do not seem to exist for frequencies above 1.0075. The second region occurs for the $+k_x$ branch near $\omega/\Omega=.998$. Here γ becomes very small and the brewster mode is ill-defined for frequencies between .998 and the top of the lower bulk band.

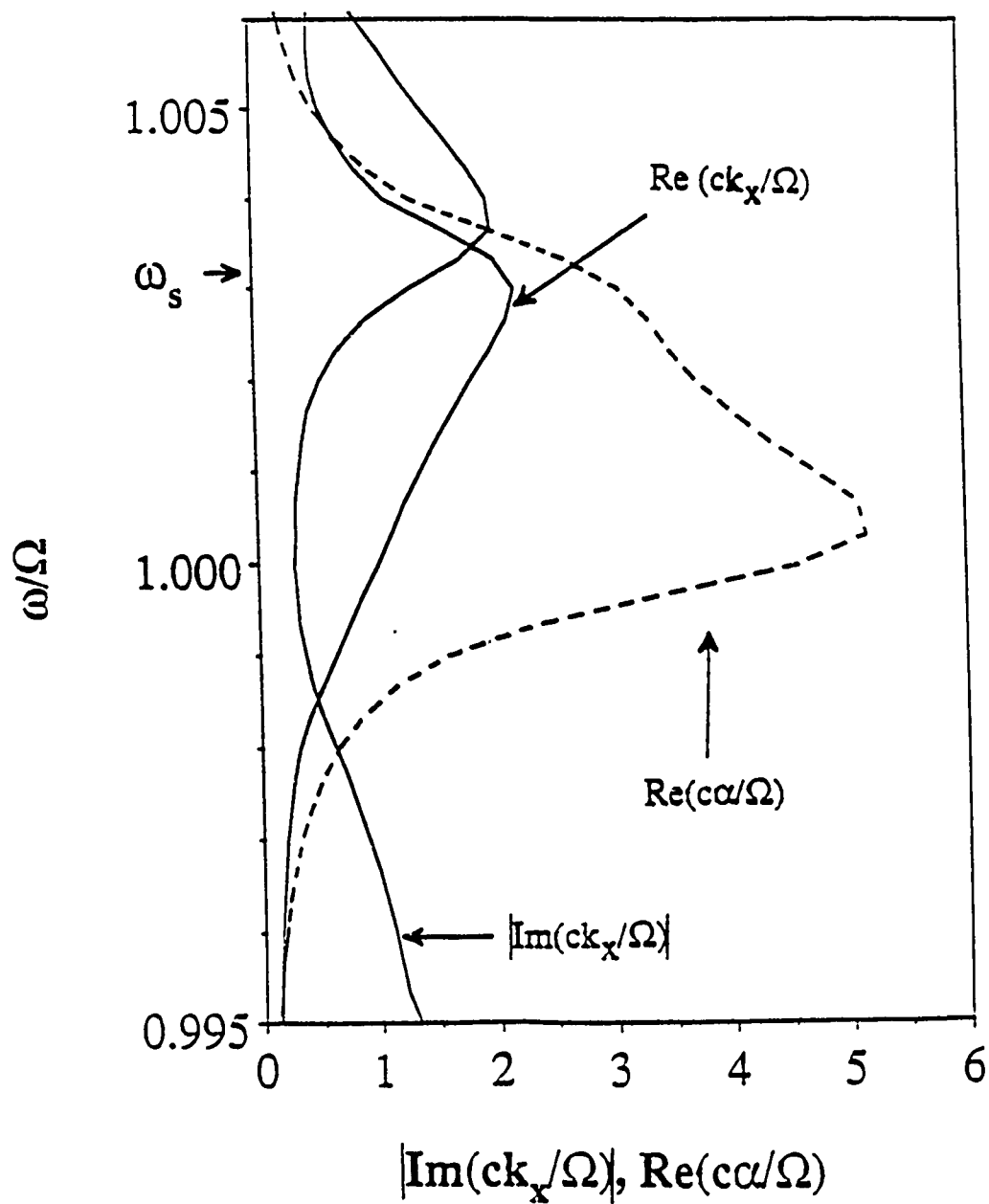


Figure 4.4. Decay parameter $|\text{Im}(ck_x/\Omega)|$ and localization parameter $\text{Re}(c\alpha/\Omega)$ for the $+k_x$ branch of dispersion curves with $1/\Omega\tau=0.0008$. The solid line is $\text{Im}(k_x)$ and the dashed line is $\text{Re}(c\alpha/\Omega)$. There is no applied field and only the magnitudes are shown. Here the increased damping decreases the localization of the modes to the surface at all frequencies.

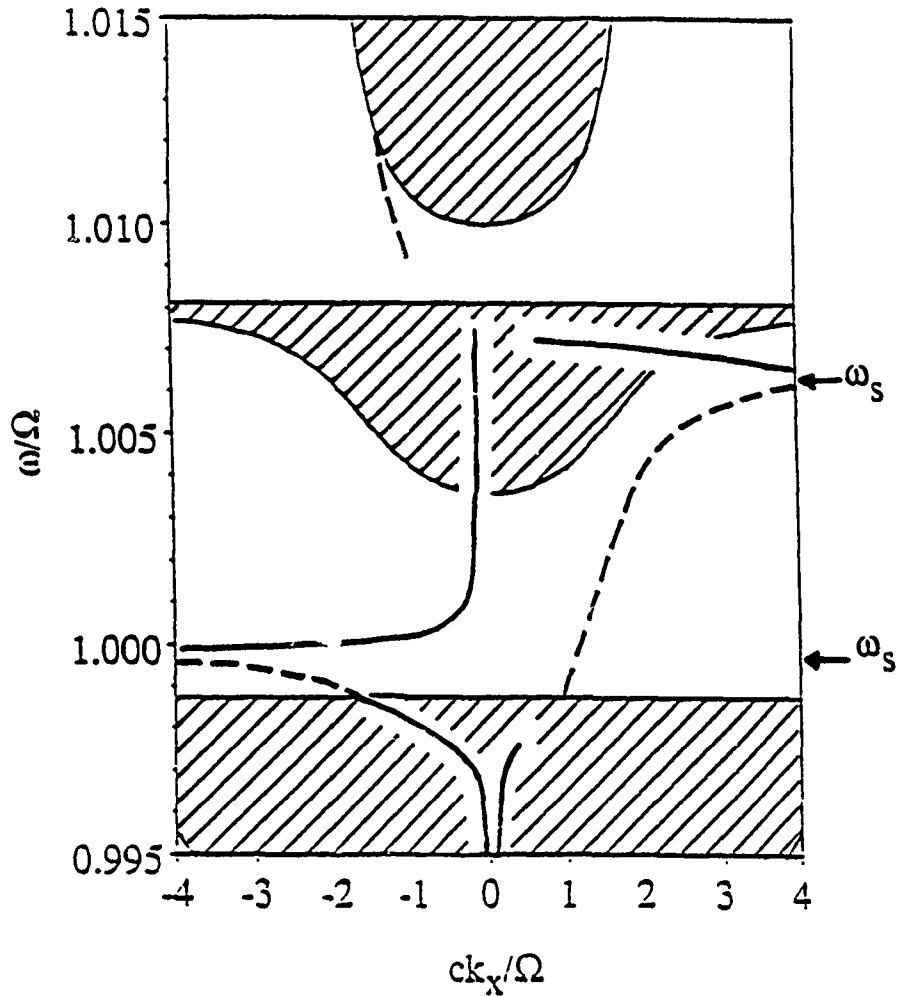


Figure 4.5. Dispersion curves for antiferromagnetic polaritons, in MnF_2 , with an applied field of .3kG. The shaded areas are bulk bands and the dashed lines are the surface modes when there is no damping present in the material. The solid lines show how damping modifies the surface mode frequencies. These solid lines are the real parts of the complex solutions to the dispersion relation (4.1) with damping $1/\Omega\tau=0.0002$. Note that the modes with damping present are reciprocal at low frequencies in the lower bulk band and become highly nonreciprocal at higher frequencies.

In figure 4.6, frequency versus $|\text{Im}(k_x)|$ and $\text{Re}(\alpha)$ are again shown, this time for the case of an applied field of 0.3 kG. The solid lines are $|\text{Im}(k_x)|$ and the dashed lines are $\text{Re}(\alpha/\Omega)$. Values for the $-k_x$ branch are shown in the top plot and values for the $+k_x$ branch are shown in the lower plot. Although signs are not shown, $\text{Im}(k_x)$ is negative for the $-k_x$ branch so that the wave attenuates in the $-x$ direction. Again there is strong localization to the surface in regions below ω_s and above the bulk band for both branches. Also note the strong localization of the surface modes where they begin near the light line (see figure 4.5) and the lack of localization as they enter the upper bulk band. Both branches are strongly bound to the surface near the lower bulk band limit and near ω_s . In between, however, the localization decreases. In both the $+k_x$ and $-k_x$ branches, the path length is large except in the lower bulk band and above ω_s . The path length is also large above the middle bulk band.

The power flows in surface excitations with damping present help clarify the nature of the leaky modes. Using the results of Camley and Mills⁶, it is a simple matter to calculate the amplitudes of the electric and magnetic fields of the surface polaritons inside and outside the material. From these one can calculate the Poynting's vector inside and outside the material. The total power flow parallel to the surface in the material is found by integrating the material Poynting's vector over a rectangular surface of width L_z from $y=0$ to $y=\infty$. Likewise, the total power flow parallel to the surface in the vacuum is found by integrating the vacuum Poynting's vector over a rectangular surface of width L_z from $y=0$ to $y=-\infty$.

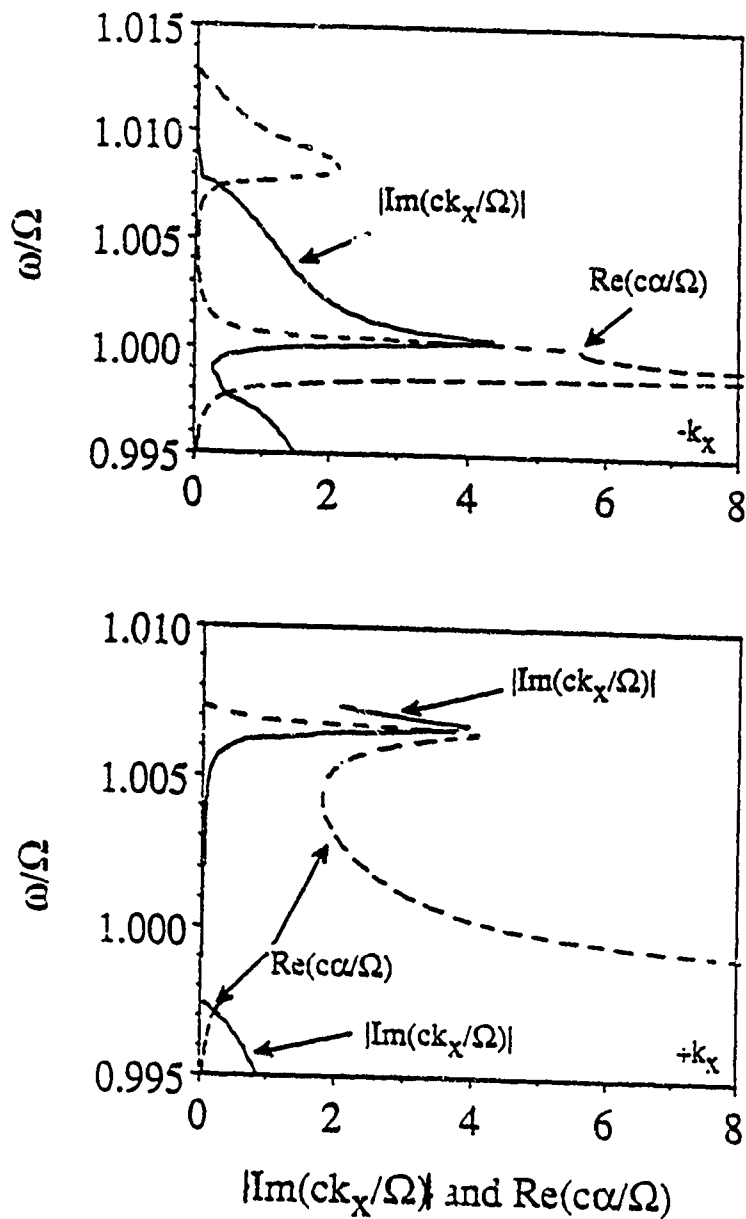


Figure 4.6. Decay parameter $|\text{Im}(ck_x/\Omega)|$ and localization parameter $\text{Re}(c\alpha/\Omega)$ for the $+k_x$ and $-k_x$ branches of the $1/\Omega\tau=0.0002$ dispersion curves of figure 4.5. Again, only the magnitudes are shown. The solid lines are $|\text{Im}(ck_x/\Omega)|$ and the dashed lines are $\text{Re}(c\alpha/\Omega)$. The decay parameters for the $-k_x$ branch are shown in the upper plot and those of the $+k_x$ branch in the lower plot. Note again the strong localization to the surface just above the lowest bulk band limit.

Defining power flows per unit length as $U^>/L_z$ for the power flow in the material and $U^</L_z$ for the power flow in the vacuum, one obtains the expressions:

$$U^< = \frac{1}{2\text{Re}(\gamma)} \text{Re} \left\{ \frac{c}{4\pi\omega_0} E_z^2 k_x \right\} \quad (4.5)$$

and

$$U^> = \frac{1}{2\text{Re}(\alpha)} \text{Re} \left\{ \frac{c}{4\pi\omega_0} E_z^2 \left(\frac{\mu_2 \alpha^* - \mu_1 k_x}{\mu_1 \alpha^* - \mu_2 k_x} \right) \gamma^* \right\} \quad (4.6)$$

Here E_z is the amplitude of the polaritons' electric field at the surface.

The parallel power flows $U^>$ and $U^<$ are plotted in figure 4.7 for the frequency and complex wavevector solutions of (4.1) for $+k_x$ with no applied field and $1/\Omega\tau = .0002$. In the bulk band ($\omega/\Omega < 1$) the power flows inside and outside the material are both in the direction of propagation. At lower frequencies, most of the energy is carried by fields in the material. Near the bulk band limit, however, most of the energy is carried in the fields outside the material, and at the antiferromagnetic resonance frequency, $\omega/\Omega = 1$, all of the electromagnetic energy is carried in the vacuum.

In the surface mode region above the bulk band and below ω_s , the energy in the material flows opposite to the direction of propagation. This is typical of surface plasmon and magnon polaritons. Note that in the surface wave region, most of the energy is carried by the fields in vacuum and so the net energy flow is in the direction of propagation except near the magnetostatic frequency. At ω_s the electromagnetic energy carried by the fields in vacuum is nearly equal to the energy carried in the opposite direction by the fields in the material and the net energy flow is very small. Above ω_s , the energy flow is in the direction of propagation both in the material and outside.

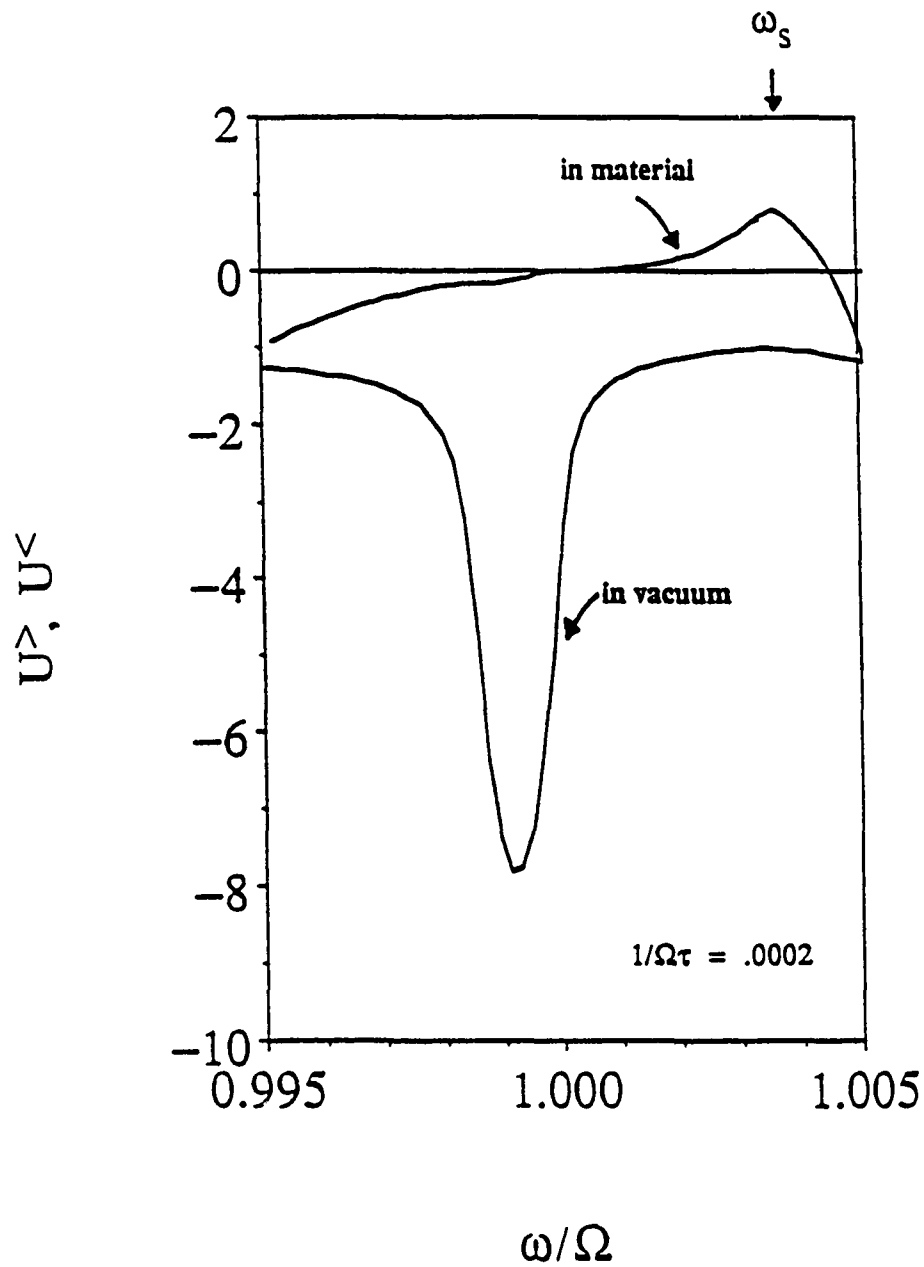


Figure 4.7. Parallel power flows $U^>$ and $U^<$ inside and outside the material for the $+k_x$ branch of $1/\Omega\tau=.0002$ modes in zero applied field. Note that in the surface mode region between ω_s and the lower bulk band limit ($\omega/\Omega=1$) the energy flow inside the material is opposite the direction of propagation.

The power flows $U^>$ and U^- are plotted in figure 4.8 for the frequency and complex wavevector solutions of (4.1) with an applied field $H_0 = 3\text{kG}$. The damping is $1/\Omega\tau = .0002$ in the upper plot and .0005 in the lower plot. Here only the $-k_x$ solutions are shown. For both dampings energy flow in the material is opposite the direction of propagation for frequencies near ω_g . Near the bulk band, however, the energy flow in the material is in the direction of propagation. Furthermore, comparison of the upper and lower plots shows that increasing the damping decreases the frequency range for energy flow in the material that opposes the direction of propagation.

The direction of energy flow in the material is governed by the sign and magnitude of the magnetic susceptibilities. A negative μ_1 usually leads to energy flows opposite the direction of propagation. This occurs in the surface mode region when there is no applied field. Also with no applied field, μ_1 is positive in the bulk band and the energy flow is in the direction of propagation. In an applied field, however, the non-vanishing μ_2 can lead to positive energy flows in the surface region and negative energy flows in the bulk, near the bulk band limits. This is because near the limit frequencies $\Omega \pm \gamma H_0$, α , μ_1 and μ_2 are extremely sensitive to frequency. Consequently, near Ω the direction of the power flow is very sensitive to the introduction of an applied field. In addition, damping also plays an important role near the bulk band limits, as seen by comparing the two plots of figure 4.8.

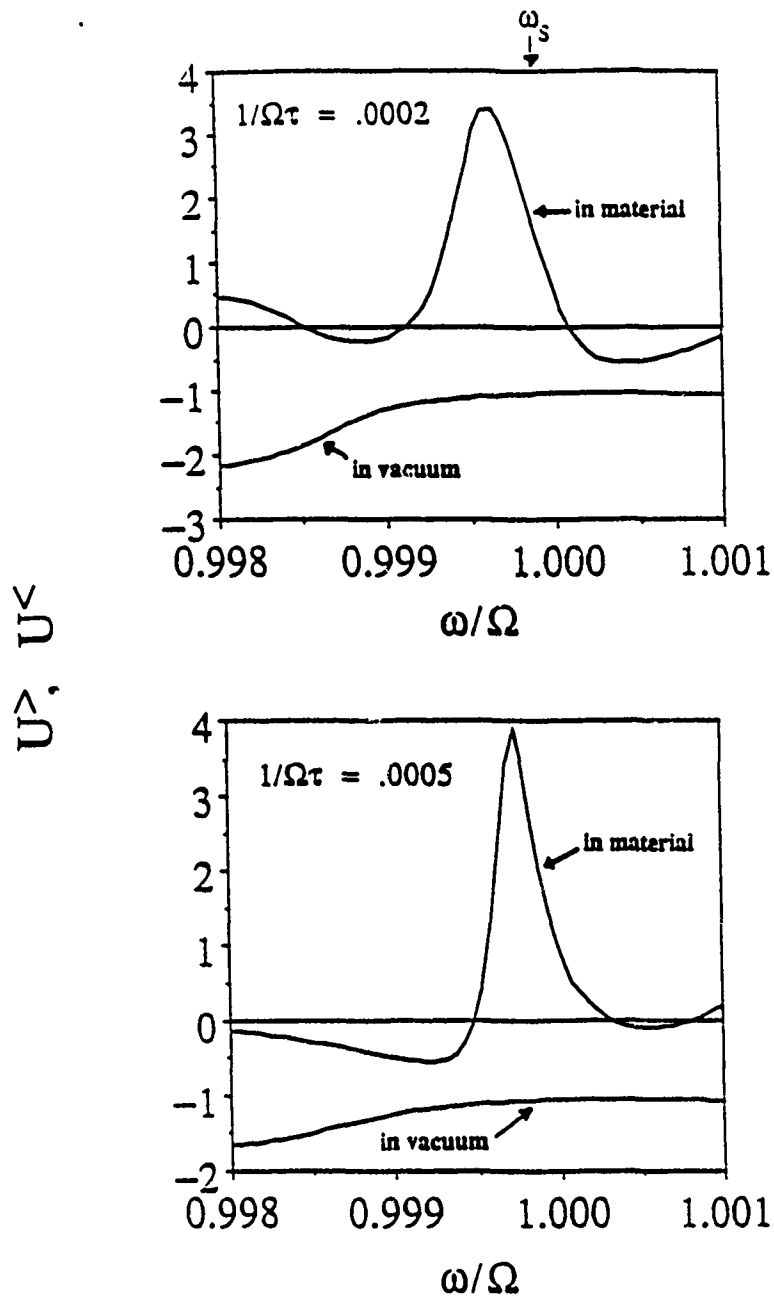


Figure 4.8. Parallel power flows $U^>$ and $U^<$ inside and outside the material for the $-k_x$ branches of $1/\Omega\tau = .0002$ modes with a .3kG applied field. The upper plot is for the case of $1/\Omega\tau = .0002$ and the lower plot is for the case $1/\Omega\tau = .0005$. Note that the direction of power flow inside the material is strongly influenced through the nonzero μ_2 term.

4.3 Surface polaritons with Landau-Lifshitz damping.

The form of the damping terms appearing in the susceptibilities originate from adding Bloch-Bloembergen terms to the equations of motion. These have the simple form

$$-\vec{M} \times \vec{M} / \tau \quad (4.7)$$

and so represent a relaxation of the magnetizations to their equilibrium values. As such, these terms can only be valid at high frequencies where the dissipative effects are averaged over several precessions of the magnetizations.

Considering the importance of damping on the nature of the leaky modes, it is of interest to examine the leaky modes when the damping mechanisms involve much shorter relaxation times. This can be accomplished by choosing damping terms of the Landau-Lifshitz form:¹⁴

$$-\Lambda \vec{M} \times \dot{\vec{M}} \quad (4.8)$$

Λ is a parameter which determines the magnitude of the damping. These terms represent a torque in the direction of the instantaneous magnetization and thus represent relaxation to the instantaneous value of the internal field. These terms are a phenomenological description which are used extensively to explain certain phenomena in ferromagnetic resonance. A derivation of the terms in (4.8) for antiferromagnets, done originally for Landau damping terms in ferromagnetic resonance,²³ can be obtained from the free energy for the magnetic system. This calculation is outlined in appendix III for the corresponding antiferromagnetic terms.

Bloch's equations of motion for the A sublattice magnetization, with Landau damping included, are

$$\dot{\vec{M}}_a = \gamma \vec{M}_a \times \vec{H}_a - \Lambda [\vec{M}_a \times (\vec{M}_a \times \vec{H}_a)] / M^2 \quad (4.9)$$

where the effective fields \vec{H}_a are defined by equations (2.3), γ is the gyromagnetic ratio and M is the saturation magnetization. A similar expression holds for the B sublattice magnetization \vec{M}_b . These constitute a set of six coupled equations which are to be put in the form

$$\vec{M} = \vec{\chi} \vec{h} \quad (4.10)$$

A time dependence $e^{i\omega t}$ is assumed and the equations are linearized in the m_x, m_y, h_x, h_y variables. This means the susceptibilities are valid only for small angles of precession for the magnetizations such that $M_z = 0$.

After straightforward algebra, the susceptibility tensor is found to have the same form as before (equation (2.3)) but the components are now given by

$$\chi_{xx} = \frac{2M\gamma[\Omega_a(\Omega_a^2 - \Omega_0^2 - \omega^2) + i\Lambda\omega(\omega^2 - \Omega_a^2 + \Omega_0^2 - 2\Omega^2)]}{\omega^4 - 2\omega^2(\Omega_0^2 + \Omega_a^2) + (\Omega_0^2 - \Omega_a^2)^2 + 2i\Lambda\omega(\omega^2 + \Omega_0^2 - \Omega_a^2)(\Omega_a^2 + \Omega^2)/\Omega_a} \quad (4.11)$$

and

$$\chi_{xy} = \frac{4i\gamma\omega\Omega_0(i\Lambda\omega - \Omega_a)}{\omega^4 - 2\omega^2(\Omega_0^2 + \Omega^2) + (\Omega_0^2 - \Omega^2)^2 + 2i\Lambda\omega(\omega^2 + \Omega_0^2 - \Omega^2)(\Omega_a^2 + \Omega^2)/\Omega_a} \quad (4.12)$$

The various quantities are $\Omega_0 = \gamma H_0$, $\Omega_a = \gamma H_a$ and $\Omega^2 = \gamma^2 H_a(H_a + 2H_e)$ where H_a is the anisotropy field, H_e is the mean exchange field, Ω is the antiferromagnetic resonance field and γ is the gyromagnetic ratio.

The remaining components are given by the simple relations $\chi_{yy} = \chi_{xx}$ and $\chi_{yx} = -\chi_{xy}$. Note that with $\Lambda=0$, these reduce to the expressions given by equations (2.4) and (2.5) for $\tau=\infty$.

The fundamental difference between these susceptibilities and the Bloch damped susceptibilities is the dependance of the damping on frequency. Since the magnitude of the damping effects increase with frequency with this case, the surface leaky modes should be most effected at higher frequencies. In figure 4.9 the real and imaginary parts of k_x and the real part of α are plotted against frequency for the case $\Lambda=.002$ and $H_0=.3\text{kG}$. The $-k_x$ branch of the dispersion curve is shown.

The general characteristics of the leaky modes are the same as in the Bloch damping case. In the Brewster mode region at lower frequencies, the leaky wave is strongly attenuated in the propagation direction and weakly bound to the surface. In the surface polariton region, the wave has a long path length and is strongly bound to the surface. In the evanescent region, where the new Landau damping effects should be most pronounced, there is again a weak localization to the surface and a shorter path length.

An interesting consequence of the Landau form of the damping is in the stability of the leaky modes in regions where the localization of the leaky modes to the surface is weak. As discussed before, when the real parts of either γ or α become vanishingly small the leaky mode wave does not represent a stable solution. These unstable frequency regions were found to be near the top of the lower and middle bulk polariton bands. When the Landau form of the damping is assumed, these frequency regions become more stable.

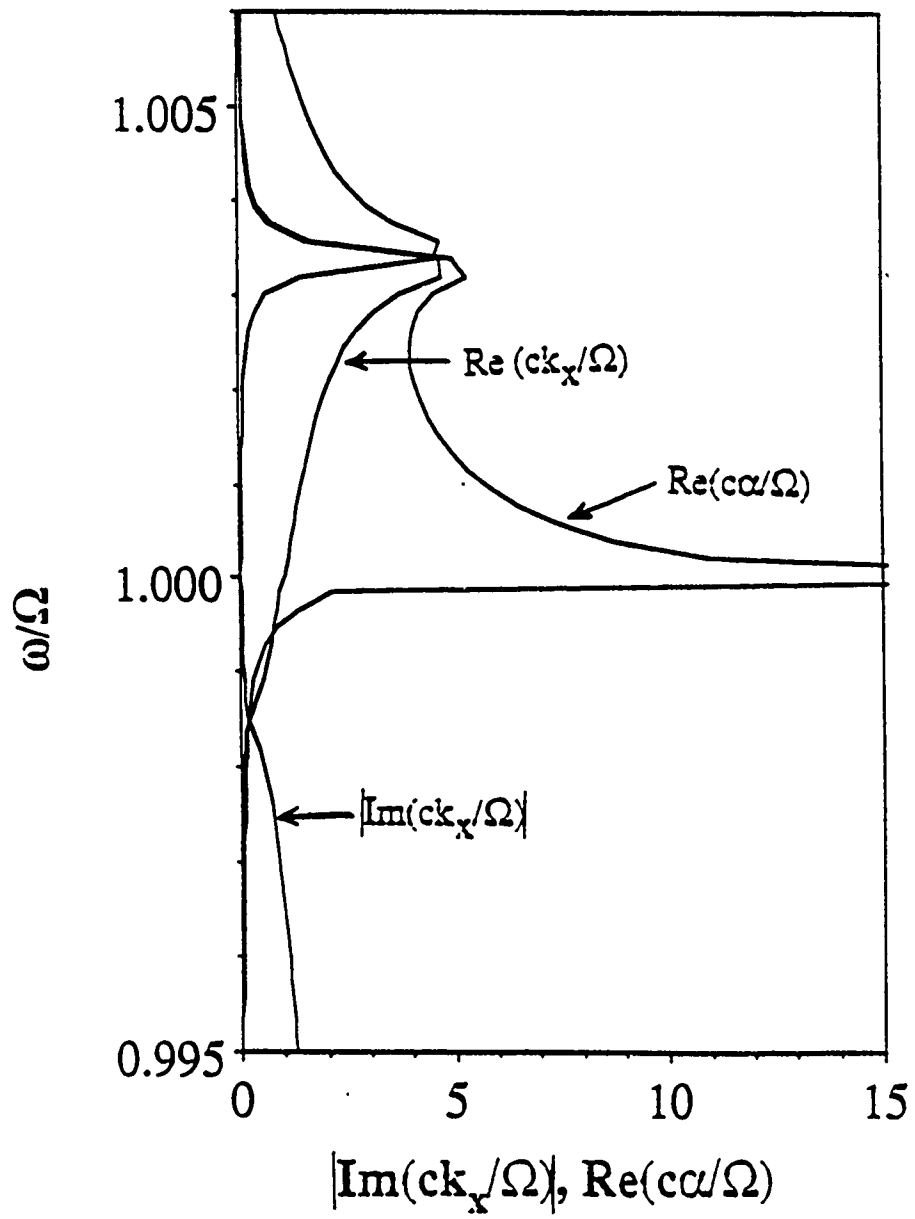


Figure 4.9. Dispersion curves and decay parameters for surface excitations on MnF_2 with no applied field. The damping terms are of the Landau form with $\Lambda=0.002$.

CHAPTER 5

ELECTROMAGNETIC GREEN'S FUNCTIONS

To date, theoretical discussions of the properties of antiferromagnetic polaritons have focused on dispersion relations calculated from the electromagnetic wave equation for anisotropic magnetic media.¹⁶ Dispersion curves describe the possible allowed modes of the magnetic system, but do not describe how these modes may be excited. It is often useful, however, to know the response of the magnetic system to some external probe-- such as the response of an antiferromagnet to an incident light wave. Green's functions are a primary tool for dealing with these types of questions.²⁴

The objective of this chapter is to calculate the electromagnetic green's functions for a semi-infinite antiferromagnet. As usual, Green's functions can be used to obtain a variety of information about the allowed modes of a system. Dispersion relations, spectral densities and the effective lifetime or attenuation length can all be obtained from the Green's functions.²⁵ In addition, Green's functions can be used in a variety of applications. Examples include the calculation of Raman and Brillouin scattering spectra, nonlinear interactions of surface polaritons, and thermodynamic properties.²⁶⁻²⁸ In the next chapter these Green's functions are used to study the reflectivity from an antiferromagnet with a slightly rough surface.

This chapter begins with the calculation of the Green's functions appropriate to a semi-infinite antiferromagnet. To simplify the algebra, only the case of polarization parallel to the easy axis is considered. Finally, the surface polariton modes discussed in previous chapters are shown to also be represented by the Green's functions.

5.1 Green's functions.

The geometry is the same as in the previous chapters. The material is in the $y > 0$ half-space with the surface at $y=0$. The antiferromagnet is uniaxial and the magnetizations of the two sublattices lie along the $\pm z$ axis. The applied field also lies along z . Inside the material, the susceptibilities are given by equations (2.3) and (2.9). Outside the material, where $y < 0$, the susceptibilities are uniform and describe a vacuum:

$$\epsilon_{ij} = \delta_{ij} \quad (5.1)$$

$$\mu_{ij} = \delta_{ij} \quad (5.2)$$

magnetic wave equation for the propagation of transverse electromagnetic waves in media described by the susceptibilities (2.3) and (2.8) under the influence of some driving field F has the form

$$\epsilon_k \sum_{j,m}' \left\{ \epsilon_m^{-1} \frac{\partial^2}{\partial x_j \partial x_k} - \delta_{jk} \left(\sum_l \epsilon_m^{-1} \frac{\partial^2}{\partial x_l^2} \right) - \omega_o^2 \mu_{kj} \right\} H_j = F_k \quad (5.3)$$

The prime on the sum means that $m \neq j, k, l$. Also, the shorthand notation $\omega_o = \omega/c$ is used. Since the dielectric tensor is diagonal in this problem, $\epsilon_{kk} = \epsilon_k$. Note that both the magnetic fields H and the driving fields F are assumed to have time dependance $e^{-i\omega t}$.

The Green's functions satisfy the associated wave equation given by

$$\epsilon_k \sum_{m,j}' \left\{ \epsilon_m^{-1} \frac{\partial^2}{\partial x_j \partial x_k} - \delta_{jk} \left(\sum_l \epsilon_m^{-1} \frac{\partial^2}{\partial x_l^2} \right) - \omega_o^2 \mu_{kj} \right\} g_{jm}(\vec{x}, \vec{x}') = 4\pi \delta_{km} \delta(\vec{x} - \vec{x}') \quad (5.4)$$

The solutions to (5.3) are then given by

$$H_j = \sum_k \int d^3\vec{x}' g_{jk}(\vec{x}, \vec{x}') F_k \quad (5.5)$$

The volume of integration is over all of space.

Equation (5.4) must be translationally invariant in both x and z . This invariance can be exploited through the use of Fourier expansions:

$$g_{jk}(\vec{x}, \vec{x}') = \int_{-\infty}^{\infty} \frac{d^2\vec{k}_{\parallel}}{4\pi^2} e^{i\vec{k}_{\parallel} \cdot (\vec{x} - \vec{x}')} g_{jk}(\vec{k}_{\parallel}; y, y') \quad (5.6)$$

and

$$\delta(\vec{x} - \vec{x}') = \delta(y - y') \int_{-\infty}^{\infty} \frac{d^2\vec{k}_{\parallel}}{4\pi^2} e^{i\vec{k}_{\parallel} \cdot (\vec{x} - \vec{x}')} \quad (5.7)$$

Here we've made the definitions $\vec{k}_{\parallel} = \hat{x}k_x + \hat{z}k_z$ for the wavevector parallel to the plane $y=0$ and $\vec{x}_{\parallel} = \hat{x}x + \hat{z}z$ for the position vector in that plane.

Applying these transformations to the Green's function equation (5.4), and writing the result in explicit matrix form,

$$\begin{bmatrix} \frac{\epsilon_1}{\epsilon_2} D^2 - k_1^2 & i(\omega^2 \mu_2 \epsilon_1 - \frac{\epsilon_1}{\epsilon_2} k_x D) & k_x k_z \\ -i(\omega^2 \epsilon_1 \mu_2 + \frac{\epsilon_1}{\epsilon_2} k_x D) & (\frac{\epsilon_1}{\epsilon_2} k_x^2 + k_1^2) & -ik_z D \\ k_x k_z & -ik_z D & D^2 - k_2^2 \end{bmatrix} \vec{g} = -4\pi \vec{I} \delta(y-y') \quad (5.8)$$

with the definitions

$$D = \partial_x \partial_y \quad (5.9)$$

$$k_1^2 = k_z^2 - \omega_0^2 \epsilon_1 \mu_1 \quad (5.10)$$

$$k_2^2 = k_x^2 - \omega_0^2 \epsilon_2 \quad (5.11)$$

The electric and magnetic fields must satisfy the usual homogeneous electromagnetic boundary conditions at the surface $y = 0$ (assuming that the sources of the driving fields are not located at the surface $y = 0$ so that the boundary conditions are homogeneous). The condition that the normal component of \vec{B} be continuous across $y = 0$ is first applied. Using the \vec{H} field given by equation (5.5) to calculate \vec{B} , applying the boundary condition and equating coefficients of F_k results in the condition

$$[-i\mu_2 g_{xm} + \mu_1 g_{ym}]_{y=0+} = [g_{ym}]_{y=0-} \quad (5.12)$$

Similarly, continuity of tangential \vec{H} results in

$$[g_{xm}]_{y=0+} = [g_{xm}]_{y=0-} \quad (5.13)$$

$$[g_{zm}]_{y=0+} = [g_{zm}]_{y=0-} \quad (5.14)$$

Finally, relating \vec{E} to \vec{H} via Maxwell's relations, continuity of tangential \vec{E} requires that

$$[i D g_{zm} + k_z g_{ym}]_{y=0+} = \epsilon_2 [i D g_{zm} + k_z g_{ym}]_{y=0-} \quad (5.15)$$

$$[k_x g_{ym} + i D g_{xm}]_{y=0+} = \epsilon_2 [k_x g_{ym} + i D g_{xm}]_{y=0-} \quad (5.16)$$

linear combination of equations (5.13) and (5.14) gives an equation identical to that which arises by requiring continuity of normal \vec{D} across $y=0$, thus providing no new information.

The complexity of the algebra makes an analytical solution for the case of general propagation directions difficult. The main features of interest, however, can be had by considering the simpler case where the electric field is constrained to lie along the z axis and the incident electromagnetic wave propagates in the xy plane. This is called the Voigt geometry and is a common experimental set up. In addition, the nonreciprocity is often a maximum in this geometry. The Voigt geometry uncouples the H_z field from the H_x and H_y fields, as can be seen by examining the equation of motion matrix (5.8). This greatly facilitates the separation and solution of the differential equations of motion and the application of the boundary conditions.

Note that in the special case of the electric field along z , g has no z dependence and thus its Fourier expansion is in one dimension only. Accordingly, the wave equation (5.4) is transformed with the one dimensional counterparts of the transforms (5.6) and (5.7).

The Green's functions differ depending on whether the source point y' lies in the material or outside in vacuum. Beginning with the case where y' lies outside the material, the matrix equation for g becomes

$$\begin{bmatrix} D^2 + \omega_o^2 \epsilon_2 \mu_1 & i(\omega_o^2 \epsilon_2 \mu_2 - k_x D) \\ -i(\omega_o^2 \epsilon_2 \mu_2 + k_x D) & -(k_x^2 - \omega_o^2 \epsilon_2 \mu_1) \end{bmatrix} \vec{g}(k_y; y, y') = 0 \quad (5.17)$$

for $y > 0$ and $y' < 0$ (source outside the material and observation point inside the material). In this polarization there is no z component of the \vec{H} field so \vec{g} is now a 2×2 matrix. The homogeneous set of equations (5.17) can be shown to have the solutions

$$g_{ij}^h = C_{ij}^> e^{-\alpha y} \quad (5.18)$$

where α represents the decay constant in the y direction and is given by

$$\alpha = + \sqrt{k_x^2 - \omega_o^2 \left(\frac{\mu_1^2 - \mu_2^2}{\mu_1} \right) \epsilon_2} \quad (5.19)$$

Note that the positive root is specified to insure proper decay in the y direction. The coefficients $C_{ij}^>$ in (5.18) will be determined later through the boundary conditions.

For fields outside the material, the appropriate Green's functions satisfy the inhomogeneous equations

$$\begin{bmatrix} D^2 + \omega_o^2 & -ik_x D \\ -ik_x D & \omega_o^2 - k_x^2 \end{bmatrix} \vec{g}(k_x, y, y') = -4\pi \vec{I} \delta(y - y') \quad (5.20)$$

for $y < 0$ and $y' < 0$.

The solutions to this set of equations will be a linear combination of a particular solution, which solves the inhomogeneous equation, and a solution to the associated homogeneous equation. Thus

$$g_{ij} = g_{ij}^h + g_{ij}^p \quad (5.21)$$

where g_{ij}^h is the homogeneous solution and g_{ij}^p is the particular solution.

The particular solutions obey

$$(D^2 - \gamma^2) g_{xx}^p = - \frac{4\pi\gamma^2}{\omega_o^2} \delta(y - y') \quad (5.22)$$

$$(D^2 - \gamma^2) g_{xy}^p = - \frac{4\pi k}{\omega_o^2} D \delta(y - y') \quad (5.23)$$

Here γ is the decay constant in the y direction outside the material. It is given by

$$\gamma = +\sqrt{k_x^2 - \omega_o^2} \quad (5.24)$$

Again, the positive root is chosen so that the wave decays away from the surface.

The solutions to the equations (5.22) and (5.23) must vanish at $y = \pm\infty$ in order to represent physical waves. These solutions are then given by the relations

$$(D^2 - a^2) \frac{e^{-a|y - y'|}}{2a} = \delta(y - y') \quad (5.25)$$

and

$$(D^2 - a^2) \frac{1}{2} \text{sgn}(y - y') e^{-a|y - y'|} = D \delta(y - y') \quad (5.26)$$

An important point to notice is that as long as a is real, the solutions vanish at $y=\pm\infty$ according to $e^{-\text{Re}(a)|y-y'|}$. Thus (5.25) and (5.26) apply even if a is complex. If a becomes imaginary, however, then in order to represent an outgoing wave at both $+\infty$ and $-\infty$ the solutions must be of the form $e^{ia|y-y'|}$. This requirement leads to relations corresponding to (5.25) and (5.26) for the case $a=ib$:

$$(D^2 + b^2) \frac{e^{ib|y-y'|}}{2ib} = \delta(y-y') \quad (5.27)$$

and

$$(D^2 + b^2) \frac{1}{2} \text{sgn}(y-y') e^{ib|y-y'|} = D \delta(y-y') \quad (5.28)$$

Comparison of (5.27) and (5.28) to (5.25) and (5.26) show that solutions for the case a imaginary are obtained from the solutions for a real by letting a go to $-ia$ in the real a solutions. The reverse transformation is obtained by letting a go to ia in the solutions for a imaginary to get the solutions for a real. Rather than write two sets of green's functions for the cases γ and α real and imaginary, only the form of the solutions for both γ and α real are presented, with the stipulation that the solutions for γ and α imaginary are obtained by letting γ go to $-i\gamma$ and α go to $-i\alpha$. Note that since 0 is a branch point for both α and γ , this prescription is equivalent to choosing a sign convention across a branch cut.

When damping is present, as it will be in the next section, α and γ become complex. In order that the Green's functions represent both an outgoing wave and a vanishing exponential at $\pm\infty$, the imaginary parts of α and γ must be negative for the forms given by (5.25) and (5.26). In order for the forms in (5.27) and (5.28) to satisfy the same boundary conditions, however, α and γ must have positive imaginary parts. The transformation rule to get from the solutions of (5.25) and (5.26) to the solutions of (5.27) and (5.28) is thus generalized by letting α go to $-i\alpha^*$ and γ go to $-i\gamma^*$ with the convention that the real parts of

α and γ are positive and the imaginary parts are negative.

Whether the imaginary and real parts of either α or γ have the same or different signs is determined by the sign on τ in the equations of motion and the boundary conditions at ∞ . As a final comment, the form of the Green's functions when γ and α are complex is chosen according to whether the frequency and wavevector k_x are in the bulk or surface regions. As discussed in Chapter 2, the surface region for waves outside the material is defined as the region in the dispersion curves where $|\gamma| > \omega_0$ and the surface region for waves in the material is where $|\alpha| > |\vec{k}_x|$. The bulk region for waves outside the material is where $|\gamma| < \omega_0$ and the bulk region for waves in the material is where $|\alpha| < |\vec{k}_x|$.

The particular solutions to equations (5.22) and (5.23) are

$$g_{xx}^p = \frac{2\pi\gamma}{\omega_o^2} e^{-\gamma|y-y'|} \quad (5.29)$$

$$g_{xy}^p = -\frac{2\pi i k_x}{\omega_o^2} \text{sgn}(y-y') e^{-\gamma|y-y'|} \quad (5.30)$$

The homogeneous solutions to the equation set (5.20) are given by:

$$g_{ij}^h = C_{ij} e^{\gamma y} \quad (5.31)$$

The next task is to derive and apply boundary conditions to determine the coefficients C_{ij} . First, the homogeneous equations of motion (5.17) is examined. These provide a relationship between g_{xx} and g_{yx} valid as y approaches 0 from the positive side:

$$\left[i(\omega_o^2 \epsilon_2 \mu_2 + k_x^2 D) g_{xx} + (k_x^2 - \omega_o^2 \epsilon_2 \mu_1) g_{yx} \right]_{y=0+} = 0 \quad (5.32)$$

Similarly, the equations of motions in (5.20) provide a relationship valid when y approaches 0 from the negative:

$$\left[ik_x D g_{xx} + (k_x^2 - \omega_o^2) g_{yx} \right]_{y=0-} = 0 \quad (5.33)$$

Together with the continuity condition on normal \vec{B} from equation (5.12), these relations result in a boundary condition on g_{xx} :

$$\left[(\omega_o^2 \epsilon_2 \mu_1 - k_x^2) D g_{xx} \right]_{y=0-} = \left[-\gamma^2 (\mu_2 k_x + \mu_1 D) g_{xx} \right]_{y=0+} \quad (5.34)$$

The continuity condition of tangential \vec{H} given in (5.13) provides a second boundary condition on g_{xx} .

For $y > 0$, the solution g_{xx} is given by equation (5.18). For $y < 0$, g_{xx} also includes the particular solution (5.29) in addition to the homogeneous part (5.31). Application of the boundary conditions (5.13) and (5.34) determine the coefficients C_{xx} from the homogeneous parts of the solutions and result in the following expressions for the solutions in each region. For $y' < 0$ and $y > 0$,

$$g_{xx}^+ = \frac{4\pi\gamma}{\omega_o^2} \left\{ \frac{A}{A-B} \right\} e^{y'} e^{-\alpha y} \quad (5.35)$$

and for $y' < 0$ and $y < 0$,

$$g_{xx}^- = \frac{2\pi\gamma}{\omega_o^2} \left\{ \frac{A \operatorname{sgn}(y-y') + B}{A-B} e^{-\gamma|y+y'|} + e^{-\gamma|y-y'|} \right\} \quad (5.36)$$

The quantities A and B are defined as

$$A = \omega_o^2 \epsilon_2 \mu_1 - k_x^2 \quad (5.37)$$

$$B = \gamma(\mu_1 \alpha - \mu_2 k_x) \quad (5.38)$$

The poles of the Green's functions occur when $A-B=0$. This condition can be shown to give the antiferromagnetic surface polariton dispersion relation derived by Camley and Mills.³

The relationships between g_{xx} and g_{yx} given in the equations of motion (5.17) and (5.20) can be used to easily determine g_{yx} from equations (5.35) and (5.36). The result for $y' < 0$ and $y > 0$ is

$$g_{yx}^+ = \frac{4\pi i \gamma}{\omega_o^2} \left\{ \frac{\omega_o^2 \epsilon_2 \mu_2 - \alpha k_x}{A-B} \right\} e^{\gamma y'} e^{-\alpha y} \quad (5.39)$$

while for $y' < 0$ and $y < 0$,

$$g_{yx}^- = \frac{2\pi i k_x}{\omega_o^2} \left\{ \frac{A \operatorname{sgn}(y-y') + B}{A-B} \operatorname{sgn}(y+y') e^{-\gamma|y+y'|} + \operatorname{sgn}(y-y') e^{-\gamma|y-y'|} \right\} \quad (5.40)$$

$$g_{xx}^- = \frac{2\pi\gamma}{\omega_o^2} \left\{ \frac{A \operatorname{sgn}(y-y') + B}{A-B} e^{-\gamma(y+y')} + e^{-\gamma|y-y'|} \right\} \quad (5.36)$$

The quantities A and B are defined as

$$A = \omega_o^2 \epsilon_2 \mu_1 - k_x^2 \quad (5.37)$$

$$B = \gamma(\mu_1 \alpha - \mu_2 k_x) \quad (5.38)$$

The poles of the Green's functions occur when $A - B = 0$. This condition can be shown to give the antiferromagnetic surface polariton dispersion relation derived by Camley and Mills.³

The relationships between g_{xx} and g_{yx} given in the equations of motion (5.17) and (5.20) can be used to easily determine g_{yx} from equations (5.35) and (5.36). The result for $y' < 0$ and $y > 0$ is

$$g_{yx}^+ = \frac{4\pi i \gamma}{\omega_o^2} \left\{ \frac{\omega_o^2 \epsilon_2 \mu_2 - \alpha k_x}{A-B} \right\} e^{\gamma y'} e^{-\alpha y} \quad (5.39)$$

while for $y' < 0$ and $y < 0$,

$$g_{yx}^- = -\frac{2\pi i k_x}{\omega_o^2} \left\{ \frac{A \operatorname{sgn}(y-y') + B}{A-B} e^{\gamma(y+y')} - \operatorname{sgn}(y-y') e^{-\gamma|y-y'|} \right\} \quad (5.40)$$

The remaining Green's functions are found in much the same manner. The particular solution for g_{xy} is given by (5.30), but boundary conditions are still required to find $C_{xy}^>$ and $C_{xy}^<$ for the homogeneous solutions given by (5.18) and (5.31). Note that in this problem the source terms do not lie at $y' = 0$, so only the homogeneous counterparts of the inhomogeneous equations belonging to (5.20) are used to uncouple the boundary conditions for g_{xy} . In particular, from (5.20) a relationship valid at $y=0^-$ is obtained:

$$\left[\gamma^2 g_{yy} + i k_x D g_{xy} \right]_{y=0^-} = 0 \quad (5.41)$$

The corresponding relation in (5.17) is:

$$\left[i (\omega_o^2 \epsilon_2 \mu_2 + k_x^2 D) g_{xy} + (\omega_o^2 \epsilon_2 \mu_1 - k_x^2) g_{yy} \right]_{y=0^+} = 0 \quad (5.42)$$

As before, this is combined with the continuity equation on normal \vec{B} (5.12) to derive the boundary condition

$$\left[-\gamma^2 (\mu_2 k_x + \mu_1 D) g_{xy} \right]_{y=0^+} = \left[(\omega_o^2 \epsilon_2 \mu_1 - k_x^2) D g_{xy} \right]_{y=0^-} \quad (5.43)$$

Again, the continuity of tangential \vec{H} (5.13) provides a second boundary condition.

Application of these conditions on the solutions given by (5.18) for $y > 0$, and the sum of (5.30) and (5.31) for $y > 0$ determine C_{xy} . The resulting expression for g_{xy} for $y' < 0$ and $y > 0$ is

$$g_{xy}^+ = - \frac{4\pi k_x}{\omega_o^2} \left\{ \frac{A}{A-B} \right\} e^{\gamma y'} e^{-\alpha y} \quad (5.44)$$

and the expression

$$g_{xy}^- = \frac{2\pi k_x}{\omega_o^2} \left\{ \frac{A + \text{sgn}(y-y')B}{A-B} e^{\gamma(y+y')} - \text{sgn}(y-y') e^{-\gamma|y-y'|} \right\} \quad (5.45)$$

for $y' < 0$ and $y < 0$.

Finally, the equations of motion in (5.17) and (5.20) are used to determine g_{yy} from the g_{xy} of equations (5.44) and (5.45). The result for $y' < 0$ and $y > 0$ is

$$g_{yy}^+ = \frac{4\pi k_x}{\omega_o^2} \left\{ \frac{\omega_o^2 \epsilon_2 \mu_2 - \alpha k_x}{A-B} \right\} e^{\gamma y'} e^{-\alpha y} \quad (5.46)$$

For $y' < 0$ and $y < 0$ the inhomogeneous term from (5.20) must be included:

$$g_{yy}^- = \frac{2\pi k_x^2}{\omega_o^2 \gamma} \left\{ \frac{A + \text{sgn}(y-y')B}{A-B} e^{\gamma(y+y')} + e^{-\gamma|y-y'|} \right\} + \frac{4\pi}{\gamma^2} \delta(y-y') \quad (5.47)$$

The above Green's functions reduce to those of Mills and Maradudin²⁶ in the nonmagnetic limit $\mu_1=1$, $\mu_2=0$, and $\epsilon_1=\epsilon_2$

5.2 Surface polaritons.

When the g_{ij} are considered as functions of k_x for an ω fixed in the surface polariton region, the g_{ij} have sharp peaks at those k_x that satisfy the surface polariton relation. To illustrate this, in figure 5.1 the imaginary part of g_{xx} (evaluated at the surface just inside the material) is plotted vs the unitless wavevector ck_x/Ω for $\omega/\Omega=1.002$ and $\omega/\Omega=.9989$. For simplicity, there is no applied field and a damping of $1/\Omega\tau=.0002$ so that the peaks have finite width.

The 1.002 peak occurs at a k_x where a surface polariton exists. At this frequency and wavevector, the excitation decays exponentially away from the surface with a negligibly small radiative part. The curve for $\omega/\Omega=.9989$, however, represents a different excitation in a frequency region forbidden to surface polaritons. This excitation is unlike the "true" surface polaritons in that it has a significantly large radiative part in directions normal to the surface.

As discussed in the derivation of the Green's functions, care must be taken in choosing the appropriate signs on the imaginary parts of γ and α in order to satisfy the outgoing wave boundary condition at infinity. When these conditions are strictly obeyed, the surface polariton peak at $\omega/\Omega=1.002$ exists but the peak at $\omega/\Omega=.9989$ does not. The peak at $\omega/\Omega=.9989$ was produced by relaxing the boundary conditions on exponentially decaying, outgoing waves at $-\infty$.

The dependance on "non-physical" boundary conditions at ∞ is a characteristic of the leaky modes and has received a great deal of attention in the literature.²⁹⁻³¹ Leaky modes are not eigenmodes of the system, and their existence is strongly dependant on the geometry of the sample and the driving electromagnetic fields. Leaky modes are typically represented as exponentially increasing into the material, with the understanding that

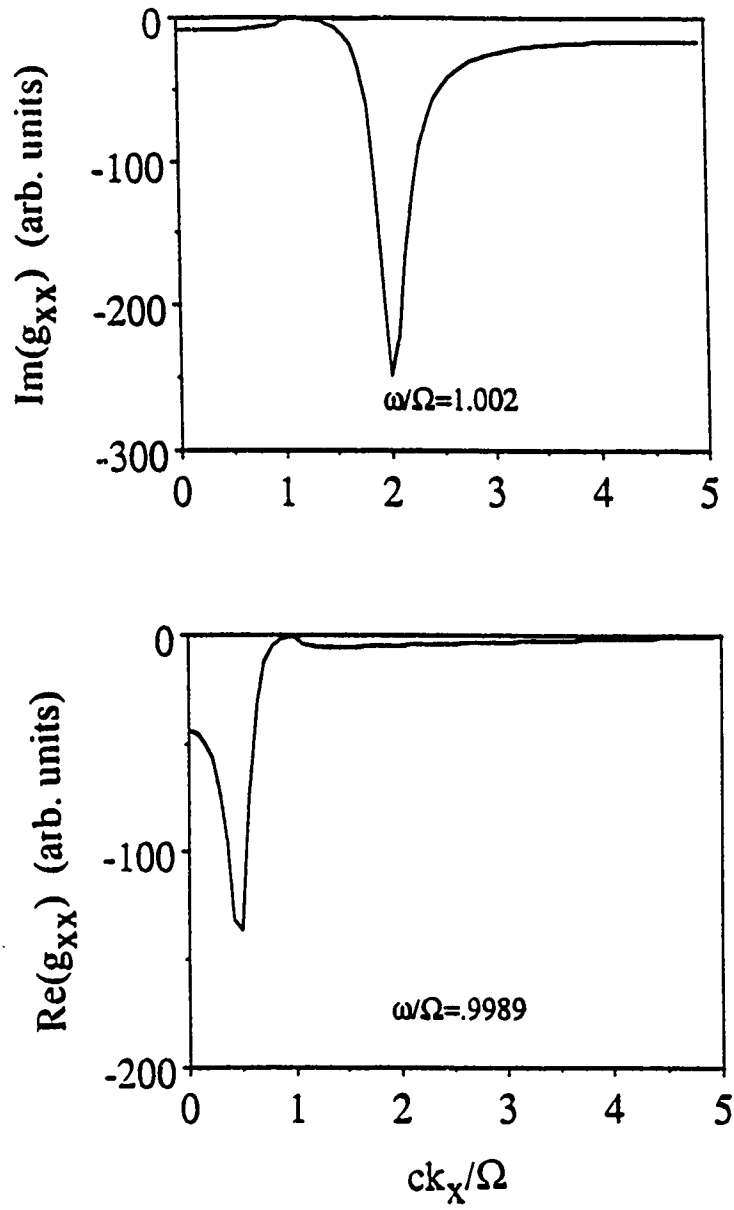


Figure 5.1. Imaginary parts of g_{xx} at $\omega/\Omega=1.002$ and $\omega/\Omega=.9989$ as functions of ck_x/Ω . There is no applied field and $1/\omega\tau=.0002$. The $\omega/\Omega=1.002$ peak occurs at the $+k_x$ surface polariton of figure 4.3. Note the existence of an excitation in the $\omega/\Omega=.9989$ curve that lies in a frequency region forbidden to surface polaritons.

absorption by the material prevents unbounded growth. The use of the Green's functions to show their existence is only an approximation and is justified mainly by the fact that leaky modes are observed in experiment and account for significant losses from the radiation fields.³¹

An approximation that ignores the boundary conditions at ∞ allows the Green's functions with damping to show excitations at real frequencies and real wavevectors that represent magnetic Brewster modes. A dispersion relation can be obtained from the Green's functions in the following way. Fix a value of ω and plot $g(k, \omega)$ as a function of k_x . The value of k_x for which there is a peak and the initial value of ω provide one point of a dispersion curve. By repeating the process for different ω values, one can trace out the entire dispersion curve.

Figure 5.2 presents the comparison between the Green's function results and the results from equation (4.1) with zero applied field. Here $1/\Omega\tau = 0.0001$, and only the region near the resonance inside the bulk band is examined. The two curves represent the results from (4.1) for ω vs. $\text{Re}(k_x)$ and the Green's function results for ω vs. k_x . While the Green's function and the $\text{Re}(k_x)$ curves are in poor agreement at smaller frequencies, they approach one another as the frequency increases. To understand this, it is useful to examine the imaginary part of the k_x as a function of frequency. It turns out that $\text{Im}(k_x)$ is large at the lower frequencies, while at higher frequencies $\text{Im}(k_x)$ is much smaller. Since the peaks in the Green's function occur at real wavevectors, there good agreement with the dispersion relation results, when k_x has a large imaginary part, is not expected.

Even despite the boundary condition approximation, the Green's functions and the dispersion curves obtained by solving (4.1) are fundamentally different things. Equation (4.1) describes possible excitations that may exist on the material, and the Green's functions

describe the propagation of waves, originating from some source, through the material. In the dispersion description, the wave attenuates according to the imaginary part of the wavevector solution. The response function, in contrast, has peaks at real frequencies and wavevectors, and it is the width of the peaks which can be related to the attenuation of the wave.

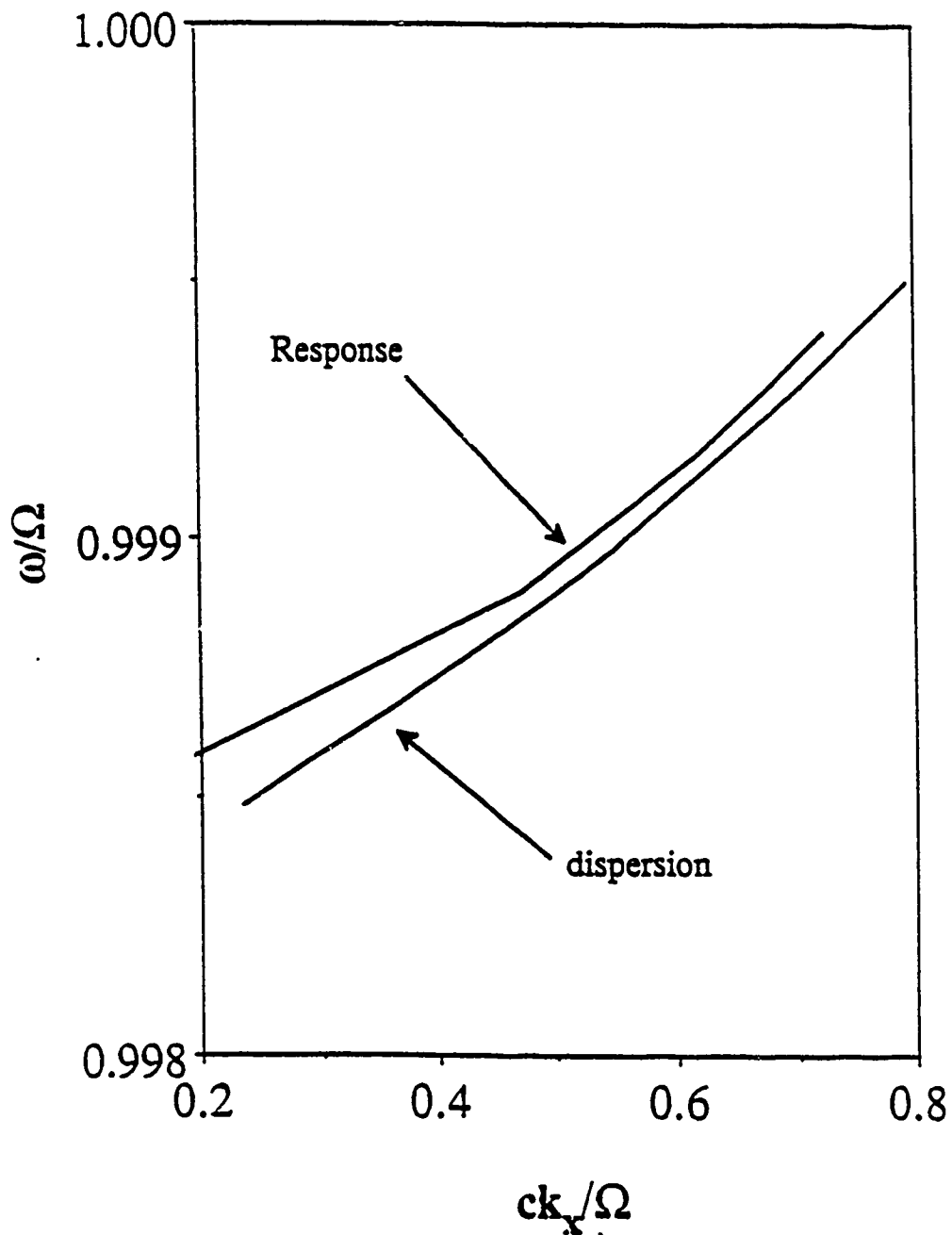


Figure 5.2. The surface resonance poles of g_{xx} as functions of real ω and real k_x with damping $1/\omega\tau=0.0001$. There is no applied field. Also plotted is the dispersion curve of the surface resonance obtained by solving the damped dispersion relation (3.1) with real frequency and complex k_x . The g_{xx} poles approach the dispersion curve at higher frequencies where $\text{Im}(ck_x/\Omega)$ becomes small.

CHAPTER 6

SCATTERING FROM ROUGH SURFACES

Numerous theoretical methods have been developed which attempt to describe the effects of various types of surface roughness. Scattering due to random roughness has been treated by perturbation methods in the limit of small surface height fluctuations using a variety of classical formulations. Earlier methods expanded the fields and surface profiles in power series under the assumption of small fluctuations in surface heights.^{32,33} Later formulations considered perturbations in the dielectric functions due to surface roughness and solved integral equations in the manner of the Born approximation of scattering theory.^{1,2} Microscopic quantum mechanical approaches have also been implemented, and the resulting expressions for scattering cross sections have been shown consistent with the classical perturbation theories.^{25,34}

The above approaches all suffer from the assumption of roughness profiles that are small with respect to the wavelength of the incident light. These perturbative techniques are usually only carried to first order and subsequently do not describe many of the more interesting features found in large amplitude studies (band gaps, frequency shifts, etc.) Alternative approaches have also been developed which attempt to deal with large amplitude gratings, and these constitute a whole subject unto themselves. Most notable are formulations of the Rayleigh hypothesis, which assumes that fields inside and outside the surface, or selvedge, region are also valid inside the selvedge region and may be used to satisfy the appropriate boundary conditions at the surface. Another technique is an integral equation method based on the "extinction theorem" which allows a formally exact solution for the scattered fields in all regions without invoking the Rayleigh assumption.^{35,36} Each

method has its disadvantages, the Rayleigh hypothesis being limited to describing fields in regions outside the surface region and the extinction theorem being difficult to implement numerically.³⁸ Consequently, new techniques are constantly evolving.

Mindful of the above remarks, the aim of this study is to provide a qualitative understanding of the effects of surface roughness on light scattering from semi-infinite uniaxial antiferromagnets. Most of the previous studies have considered the effects of roughness on only plasmon polaritons and ferromagnetic magnons.²⁷ The most interesting case is the scattering of light near antiferromagnetic surface polariton frequencies, where the scattered light's amplitude and linewidth are related to the total magnitude of the power scattered into surface polariton states.

The roughness is assumed to be slight and the perturbation method used by Mills and Maradudin in their initial studies of scattering from rough metals is applied.²⁶⁻²⁷ This method, though not without certain shortcomings, provides a reasonable first attack on the problem and can be shown to be consistent with other techniques, at least to first order in the roughness height. Since this technique requires the Green's functions for a semi-infinite antiferromagnet, the results of the last chapter are used.

Once expressions for the roughness induced scattered fields are obtained, the power losses due to roughness from a beam specularly reflected by the surface of an antiferromagnet are estimated. Since random roughness can be considered as a superposition of periodic gratings of varying heights and periods, the first case considered is for a sinusoidal one dimensional grating impressed upon the material's surface. The randomly rough surface is represented as an average of surface profiles, as also done by Mills and Maradudin, with a Gaussian distribution of surface heights and profiles.

Besides allowing an incident beam to couple with either or both the $+k$ and the $-k$ surface polariton branches, surface roughness can *enhance* the nonreciprocal reflectivity of an antiferromagnet in an applied field. In other words, in an applied field the difference in reflectivity between a beam incident at $+\theta$ and a beam incident at $-\theta$, both of which couple to

the surface polariton modes, is increased by the presence of random roughness. There is also the interesting result that roughness allows coupling between an incident beam and the Brewster-like modes of a damped antiferromagnet. The properties of these damped modes are discussed in detail in chapter 4.

The outline of the present chapter is as follows: in section 6.2 the integral equation for the scattered fields is derived and solved for the power flow inside and outside the material. In section 6.3 a small amplitude periodic grating is used to study coupling to surface polaritons on MnF_2 . Finally, in section 6.4 the case of random surface roughness is examined.

6.1 Born approximation for scattered fields.

The geometry is the same as before with the exception that the surface is no longer smooth. The actual surface height is some function $\zeta(x, z)$ which measures the deviation of the actual surface above an ideal smooth surface at $y = 0$. This rough surface geometry is shown in figure 6.1. The explicit forms of $\vec{\epsilon}$ and $\vec{\mu}$ are given by equations (2.3) and (2.8).

The electric and magnetic susceptibilities are uniform everywhere except near the surface of the material. The only x and z dependence of the susceptibilities enters through the profile function $\zeta(x, z)$. At long wavelengths, a step function behavior of the susceptibilities at the surface is appropriate, and the dependence of the susceptibilities on the roughness profile $\zeta(x, z)$ is represented by the step function $\Theta(\zeta(x, z) - y)$:

$$\epsilon_{ij}(\vec{x}) = \epsilon_{ij}^0 + (\delta_{ij} - \epsilon_{ij}^0) \Theta(\zeta(x, z) - y) \quad (6.1)$$

$$\mu_{ij}(\vec{x}) = \mu_{ij}^0 + (\delta_{ij} - \mu_{ij}^0) \Theta(\zeta(x, z) - y) \quad (6.2)$$

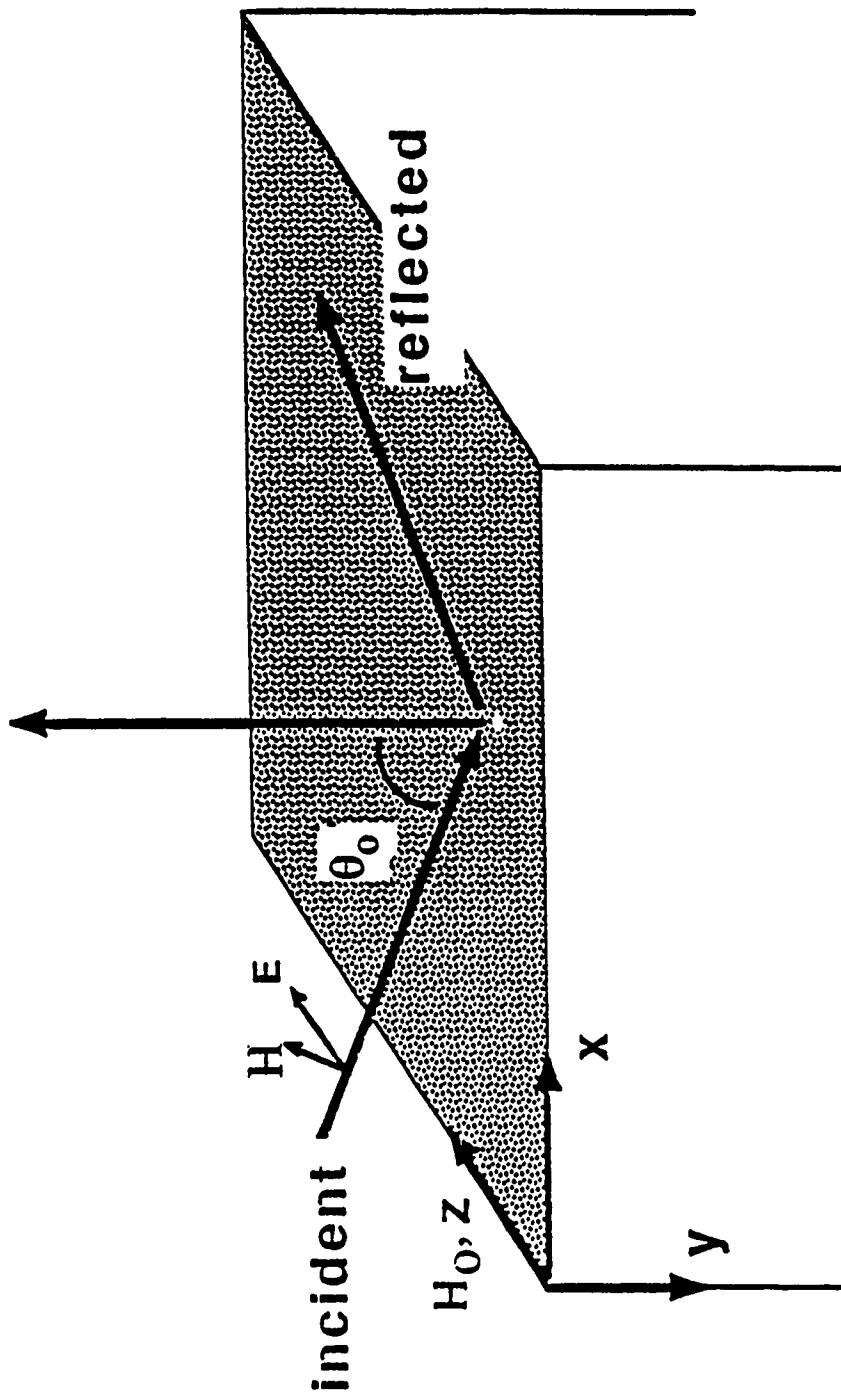


Figure 6.1. Scattering geometry. An electromagnetic wave with its electric field along z and H field in the xy plane is incident on an antiferromagnet with a rough surface. The material lies in the $y > 0$ half space with the easy axis along z . An applied field is also set along the z axis. The angle of incidence, θ_0 , is defined between the incident wavevector and the outward normal from the surface of the antiferromagnet.

Here δ_{ij} is the kronecker delta function which makes the dielectric and magnetic susceptibility tensors equal to unit matrices for the vacuum region. Note that we've written the susceptibilities as tensor components. The superscripts 'o' denote the values appropriate for the bulk material and the unsubscripted ϵ_{ij} and μ_{ij} are the position dependent functions for the entire structure with the rough surface.

The roughness is treated as a small perturbation in the bulk material. If $\widehat{\epsilon}_f$ and $\widehat{\mu}_f$ are the position dependent flat surface susceptibility tensors, then the position dependent rough surface susceptibility tensors ϵ and μ are expanded to first order in the perturbation as:

$$\widehat{\epsilon} = \widehat{\epsilon}_f + \delta\widehat{\epsilon} \quad (6.3)$$

$$\widehat{\mu} = \widehat{\mu}_f + \delta\widehat{\mu} \quad (6.4)$$

Expressions of this form can be found by expanding the susceptibilities in (2.1) and (2.2) in a Taylor's series about $\zeta(x,z)=0$. The first order terms appearing in this expansion are

$$\delta\epsilon_{ij} = \zeta(x,z) (\delta_{ij} - \epsilon_{ij}^o) \delta(-y) \quad (6.5)$$

$$\delta\mu_{ij} = \zeta(x,z) (\delta_{ij} - \mu_{ij}^o) \delta(-y) \quad (6.6)$$

The electromagnetic fields are also written in a perturbation expansion. The unperturbed fields are \widehat{H}_0 and \widehat{E}_0 and the first order corrections are \widehat{H}^{sc} and \widehat{E}^{sc} . A time dependence for \widehat{H} and \widehat{E} of the form $\exp(-i\omega t)$ is assumed. From Maxwell's equations a wave equation for the scattered fields, \widehat{H}^{sc} , is derived and is valid to first order in the perturbations:

$$\nabla \times \epsilon_0^{-1} \nabla \times \vec{H}^{sc} - \omega_0^2 \mu_0 \vec{H}^{sc} = \omega_0^2 \delta \mu \vec{H}_0 - i \omega_0 \nabla \times \epsilon_0^{-1} \delta \epsilon \vec{E}_0 \quad (6.7)$$

For the rest of the paper, $\omega_0 = \omega/c$. An equivalent expression for (6.7) is

$$\epsilon_k \sum_j \left\{ \epsilon_m^{-1} \frac{\partial^2}{\partial x_j \partial x_k} - \delta_{jk} \left(\sum_l \epsilon_m^{-1} \frac{\partial^2}{\partial x_l^2} \right) - \omega_0^2 \mu_{kj} \right\} H_j = F_k \quad (6.8)$$

The prime on the sum means that $m \neq j, k, l$. Since the dielectric tensor is diagonal, $\epsilon_{kk} = \epsilon_k$.

Note that for fields in the vacuum region the susceptibilities are given by $\mu_{ij} = \delta_{ij}$ and $\epsilon_{ij} = \delta_{ij}$.

The vector \vec{F} represents the driving terms and is given by

$$F_k = \omega_0^2 \epsilon_k \sum_j \delta \mu_{kj} H_j^0 - i \omega_0 \epsilon_k \sum_{j,l} e_{kjl} \frac{\partial}{\partial x_j} \left(\frac{\delta \epsilon_l}{\epsilon_l} E_l^0 \right) \quad (6.9)$$

Here e_{jkl} is the Levi-Civita tensor.

The perturbed wave equation (6.8) is solved by a Green's function method. The scattered field is determined by

$$H_j^{sc} = \sum_k \int d^3 \vec{x}' g_{jk}(\vec{x}, \vec{x}') F_k \quad (6.10)$$

where g_{ij} is the ij^{th} component of the semi-infinite antiferromagnetic green's function tensor.

The volume of integration is over all of space. Note that \vec{F} is a function of the unperturbed \vec{H}^0 and \vec{E}^0 fields appropriate to the material region.

For reasons discussed in chapter 5, the simpler case of propagation in the xy plane with the incident beam polarized perpendicular to the plane of incidence is considered. In this case none of the quantities in (6.8) depend on z and a Fourier transform in the x variable is used. The one dimensional Fourier expansion for g_{ij} is

$$g_{ij}(\vec{x}, \vec{x}') = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} e^{ik_x(x-x')} g_{ij}(k_x; y, y') \quad (6.11)$$

The relevant quantities for the transform of \vec{F} are \vec{H}^0 and ζ . Also, only surface irregularities that depend only on the x coordinate and are independent of z can be described in this framework. This restriction allows a reasonable description of gratings but is incapable of fully describing a randomly rough surface. To see this, suppose the roughness varied in the z direction. One would then expect scattering into waves which do not propagate perpendicularly to the applied field and in addition there could be some scattering of the perpendicular polarized incident wave into parallel polarized states. In this simplified theory, only scattering within the same polarization state is accounted for. This theory can thus describe a grating with a variable period (one dimensional random roughness), but not a true two dimensional random surface.

The x dependent surface profile function has the one dimensional expansion

$$\zeta(x) = \int_{-\infty}^{\infty} \frac{dk_x}{2\pi} e^{-ik_x x} \zeta(k_x) \quad (6.12)$$

Finally, the unperturbed fields have only one Fourier component:

$$H_i^o(\mathbf{x}) = e^{ik_{x0}x} H_i^o(k_{x0}, y') \quad (6.13)$$

$$E_z^o(\mathbf{x}) = e^{ik_{x0}x} E_z^o(k_{x0}, y') \quad (6.14)$$

k_{x0} is the wavevector of the unperturbed wave. In this problem, the unperturbed field originates from a travelling electromagnetic wave incident on the rough surface of the antiferromagnet. The amplitude of the unperturbed fields in (6.13) and (6.14) then depends on k_{x0} and is determined in each region of space by the appropriate Fresnel relations. The pertinent formulas are presented in the Appendix.

Using the transforms (6.11-6.14), the scattered field of (6.10) can be written in the compact notation

$$\vec{H}^{sg}(\vec{x}) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} dk_x \vec{\Lambda}(k_x, y) \zeta(k_x - k_{x0}) e^{ik_x x} \quad (6.15)$$

The components of the vector Λ are defined by

$$\begin{aligned} \Lambda_x = & \int_{-\infty}^{\infty} dy' \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_x^o(k_{x0}, y') - i\mu_2 H_y^o(k_{x0}, y')] + iE_z^o \frac{\omega}{\epsilon_2} (1-\epsilon_2) \frac{\partial}{\partial y} \right\} g_{xx}(k_x, y, y') \delta(-y') \\ & + \int_{-\infty}^{\infty} dy' \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_y^o(k_{x0}, y') + i\mu_2 H_x^o(k_{x0}, y')] + iE_z^o \frac{\omega}{\epsilon_2} (1-\epsilon_2) k_x \right\} g_{xy}(k_x, y, y') \delta(-y') \end{aligned} \quad (6.16)$$

and

$$\begin{aligned} \Lambda_y = & \int_{-\infty}^{\infty} dy' \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_x^o(k_{x0}, y') - i\mu_2 H_y^o(k_{x0}, y')] + iE_z^o \frac{\omega}{\epsilon_2} (1-\epsilon_2) \frac{\partial}{\partial y} \right\} g_{yx}(k_x, y, y') \delta(-y') \\ & + \int_{-\infty}^{\infty} dy' \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_y^o(k_{x0}, y') + i\mu_2 H_x^o(k_{x0}, y')] + iE_z^o \frac{\omega}{\epsilon_2} (1-\epsilon_2) k_x \right\} g_{yy}(k_x, y, y') \delta(-y') \end{aligned} \quad (6.17)$$

In deriving these formulas an integration by parts has been done, shifting the derivatives off of $\vec{\delta\epsilon\vec{E}}$ and onto the green's functions.

As noted by Mills and Maradudin²⁵ and others³⁷, the evaluation of the integrals in (6.16) and (6.17) is not obvious since the discontinuity of the integrand allows for a number of possible solutions. Following Mills, a prescription consistent with previous perturbative treatments is taken. This prescription makes the source of the scattered fields proportional to the amplitude of the unperturbed fields inside the material, as they are in equation (6.9), but

located in vacuum just outside the $y=0$ plane. Hence the Green's functions appropriate for sources in the $y < 0$ space are used with driving fields appropriate to the $y > 0$ space.

These choices result in the following expressions for Λ :

$$\begin{aligned} \Lambda_x = & \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_x^0(k_{x0}+) - i\mu_2 H_y^0(k_{x0}+)] + iE_z^0 \frac{\omega}{\epsilon_2} (1-\epsilon_2) \frac{\partial}{\partial y} \right\} g_{xx}(k_x; y, \cdot) \\ & + \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_y^0(k_{x0}+) + i\mu_2 H_x^0(k_{x0}+)] + iE_z^0 \frac{\omega}{\epsilon_2} (1-\epsilon_2) k_x \right\} g_{xy}(k_x; y, \cdot) \end{aligned} \quad (6.18)$$

and

$$\begin{aligned} \Lambda_y = & \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_x^0(k_{x0}+) - i\mu_2 H_y^0(k_{x0}+)] + iE_z^0 \frac{\omega}{\epsilon_2} (1-\epsilon_2) \frac{\partial}{\partial y} \right\} g_{yx}(k_x; y, \cdot) \\ & + \epsilon_2 \left\{ \omega_0^2 [(1-\mu_1)H_y^0(k_{x0}+) + i\mu_2 H_x^0(k_{x0}+)] + iE_z^0 \frac{\omega}{\epsilon_2} (1-\epsilon_2) k_x \right\} g_{yy}(k_x; y, \cdot) \end{aligned} \quad (6.19)$$

The fields and Green's functions are evaluated at the flat surface, the \pm signs indicating to which side of $y'=0$ the expressions belong. The derivatives, of course, are evaluated before the limiting process.

The energy flow in the scattered fields can now be calculated. The time averaged Poynting vector is

$$\vec{S} = (c/4\pi) \vec{E}^{sc} \times \vec{H}^{sc*} \quad (6.20)$$

One further notational device is required. Since the Green's functions and wavevectors for $y > 0$ are different than those for $y < 0$, the expressions \vec{E} , \vec{H} and \vec{S} will also differ depending on whether they are taken in the material or outside of it. For the rest of the paper then, the subscript ">" will indicate that an expression is valid inside the material and "<" will indicate that an expression is valid outside the material (in vacuum). To reduce the number of equations which follow, it is useful to define $\epsilon_{>} = \epsilon_0$ and $\epsilon_{<} = 1$.

Away from the rough surface, the perturbed fields satisfy

$$\omega_0 \epsilon \vec{E}^{sc} = - \vec{k} \times \vec{H}^{sc} \quad (6.21)$$

where the ">" indicates that one must choose either the expression appropriate to the $y < 0$ or the $y > 0$ region. The perturbed electric fields are

$$\vec{E}_{<}^{sc} = - \frac{1}{4\pi^2 \omega_0} \epsilon_{<}^{-1} \int_{-\infty}^{\infty} dk_x \vec{k}_{<} \times \vec{\Lambda}_{<}(k_x) \zeta(k_x - k_{xo}) e^{ik_x x} \quad (6.22)$$

The wavevector inside the material is given by

$$\vec{k}_{>} = \hat{x}k_x + \hat{y}k_{y>} \quad (6.23)$$

where the normal component, $k_{y>}$, is found through the dispersion relation

$$k_{y>} = + \sqrt{\omega_0^2 \left(\frac{\mu_1^2 - \mu_2^2}{\mu_1} \right) \epsilon_2 - k_x^2} \quad (6.24)$$

This relation is found by setting the determinant of the homogeneous equations of motions

to zero as discussed in previous chapters. The vacuum wavevector outside the material, $\vec{k}_<$, is defined by

$$\vec{k}_< = \hat{x}k_x + \hat{y}k_{y<} \quad (6.25)$$

with the normal component given by the free space dispersion

$$k_{y<} = -\sqrt{\omega_o^2 - k_x^2} \quad (6.26)$$

Note that in both (6.24) and (6.26) the signs on the radicals have been chosen so as to represent waves decaying exponentially away from the surface when the arguments of the radicals are real and negative. Likewise, the signs indicate waves travelling away from the surface when the arguments of the radicals are positive and real.

Taking the product $\vec{E}^{sc} \times \vec{H}^{sc*}$, the following quantities are defined

$$\begin{aligned} \vec{P}_{\hat{z}}(k_x, y) = & \frac{-c}{64\pi^2 \omega_o^2} \int_{-\infty}^{\infty} dk_x' \left(\vec{E}_{\hat{z}}^{-1} \vec{k}_{\hat{z}} \times \vec{\Lambda}_{\hat{z}}(k_x, y) \right) \times \vec{\Lambda}_{\hat{z}}^*(k_x', y) \\ & \times \zeta(k_x - k_{xo}) \zeta^*(k_x' - k_{xo}') e^{ix(k_x - k_x')} \end{aligned} \quad (6.27)$$

The vector functions $\vec{P}(k_x, y)$ give the scattered power flow per unit area with wavevector component parallel to the surface in a range $k_x, k_x + dk_x$, at a distance y away from the surface of the material.

The scattered power flows are then given simply by

$$\vec{S}_{\geq}(y) = \int_{-\infty}^{\infty} dk_x \vec{P}_{\geq}(k_x, y) \quad (6.28)$$

Since the scattered fields and power flows depend on y through a complex exponential term originating in the Green's functions, for further calculations it is useful to illustrate this dependence explicitly by writing

$$\vec{A}_{\geq}(k_x, y) = \vec{A}_{\geq}(k_x) e^{ik_y y} \quad (6.29)$$

A y independent portion of \vec{P} can also be defined, but the specific form depends on the profile function $\zeta(k_x)$. Two different forms for $\zeta(k_x)$ are chosen here: one for a periodic grating (see equation (6.43)), and one for a random grating (see equation (6.51)). When averaged over the profile distribution for the random grating, the integral over k_x' in equation (6.27) is equal to the product $\vec{p}_{\geq}(k_x) \exp(-2y \text{Im}(k_{y\geq}))$. $\vec{p}_{\geq}(k_x)$ is given by

$$\vec{p}_{\geq}(k_x) = \frac{ch^2 \sigma^2}{256\pi^6 \omega_0^2} e^{-k_x^2 \sigma^2 / 4} \left(\vec{\epsilon}_{\geq}^{-1} \vec{k}_{\geq}(k_x) \times \vec{\lambda}_{\geq}(k_x) \right) \times \vec{\lambda}_{\geq}^*(k_x) \quad (6.30)$$

Here h is the average height of the roughness, and σ is the correlation length of the random profile distribution.

Finally, note that it is only the real part of (6.28) that is meaningful. Consequently, in the calculations which follow, it is to be understood that only the real parts of (6.28) and (6.30) are used where appropriate.

For a grating of period s , the integral in (6.27) has a more complicated form:

$$\vec{P}_z(k_x) = \sum_q \vec{p}_z(k_x, q) e^{-2y \operatorname{Im}(k_y(q))}$$

where q takes the values $k_{x0}+s$, $k_{x0}-s$, and k_{x0} with k_{x0} the x component of the incident wavevector. When $q = k_{x0}+s$ or $k_{x0}-s$, $\vec{p}_z(k_x, q)$ is given by

$$\vec{p}_z(k_x, q) = \frac{ch^2}{64\pi^5 \omega_0^2} \delta(k_x - q) (\vec{\epsilon}_z^{-1} \vec{k}_z(q) \times \vec{\lambda}_z(q)) \times \vec{\lambda}_z^*(q)$$

Here h is the grating height. The calculation of (6.31) and (6.32) involve averages over x under the assumption that the width of the illuminated surface area is much larger than s . This is discussed in section 6.3. Finally, when $q = k_{x0}$, $\vec{p}_z(k_x, q)$ contains the factor 256 instead of the 64.

Next the question of how to represent the power lost from the specularly reflected beam is considered. First, the Poynting vectors (6.28) are normalized to the illuminated surface area by dividing \vec{p} with

$$P_o = L_x L_z (c/4\pi) H^{(i)2} \cos \theta_o \quad (6.31)$$

$H^{(i)}$ is the amplitude of the unperturbed incident field, L_x is the width in x of the area illuminated by the incident beam and L_z is the width in z . θ_o is the angle of incidence defined in figure 6.1.

The real part of the normalized expressions are integrated over an appropriate surface to obtain the net scattered energy flow per unit surface area. The vacuum energy flows are derived first. The calculation begins by separating the k_x integral into integrals over two different regions. In the region $|k_x| < |\vec{k}_<|$ the integral describes the flow of energy in radiative states.

Defining this flow as $I_{<}^r$, then

$$I_{<}^r = \int \vec{da} \cdot \int_{|k_x| < |\vec{k}_<|} dk_x \frac{\vec{p}(k_x)}{P_o} e^{2y \text{Im}(k_y)} \quad (6.32)$$

The surface of integration for the radiative integral is a cylinder of radius $L_x/2$ and length L_z whose axis coincides with the z axis of this geometry. In the vacuum region, $k_{y<}$ is pure real for the radiated energy and the integral is straightforward:

$$I_{<}^r = \frac{L_x L_z}{P_o} \int_{|k_x| < |\vec{k}_<|} dk_x \hat{y} \cdot \vec{p}_<(k_x) \quad (6.33)$$

In the region $|k_x| > |\vec{k}_<|$ the integral describes the flow of energy in evanescent states. Calling this $I_{<}^e$, then

$$I_{<}^e = \int \vec{da} \cdot \int_{|k_x| > |\vec{k}_<|} dk_x \frac{\vec{p}_<(k_x)}{P_o} e^{2y \text{Im}(k_y)} \quad (6.34)$$

The energy flow in the evanescent fields is parallel to the surface in the $\pm x$ direction, so the surface integral is over a strip of width L_z that extends from $y=0$ to $y=\infty$. The surface element da is always taken in the $+x$ direction. Since $k_{y<}$ is pure imaginary for all k_x in this

wavevector region, the surface integral is easily performed with the result

$$I_{<}^e = \frac{L_z}{P} \int_{|\vec{k}_x| > |\vec{k}_{<}|} dk_x \frac{\hat{x} \cdot \vec{p}_{<}(k_x)}{2|\text{Im}(k_{y<})|} \quad (6.35)$$

In the material, $k_{y>}$ may be complex if damping exists. For the half cylinder region within the material, the imaginary part of k_y in the radiative integral is due solely to damping and represents an exponential decay of the wave into the material. Under the assumption that the width of the illuminating beam, L_x , is much smaller than the decay length, the surface integral in the material is done exactly as in the vacuum case:

$$I_{>}^r = \frac{L_x L_z}{P} \int_{|\vec{k}_x| < |\vec{k}_{>}|} dk_x \hat{y} \cdot \vec{p}_{>}(k_x) \quad (6.36)$$

The evanescent flows in the material are found in the same manner as in the vacuum case (except with the limits of integration over y from 0 to $-\infty$) with the result

$$I_{>}^e = \frac{L_z}{P} \int_{|\vec{k}_x| > |\vec{k}_{>}|} dk_x \frac{\hat{x} \cdot \vec{p}_{>}(k_x)}{2|\text{Im}(k_{y>})|} \quad (6.37)$$

The ultimate goal is to calculate the change in reflectance due to the surface roughness. First the sums $I_{>} = |I_{>}^e| + |I_{>}^r|$ and $I_{<} = |I_{<}^e| + |I_{<}^r|$ are defined. It is important to note that while $I_{>}$ and $I_{<}$ represent scattering into states inside and outside the material respectively, they do not by themselves determine the fraction of power scattered out of the specularly reflected beam. The most one can say is that if ΔR and ΔT represent the changes in reflectance and transmittance due to roughness induced scattering, then

$$(I_{>} + I_{<})_{ns} = \Delta R + \Delta T \quad (6.38)$$

Only the nonspecular portions of the scattering ratios in (6.38) are needed to calculate the losses from the specular beams. This is emphasized by the subscript "ns".

To approximate ΔR , first write

$$\Delta R = (I_{>} + I_{<})_{ns} / (1 + \Delta T / \Delta R) \quad (6.39)$$

and assume that the perturbing roughness does not have an appreciable effect on the ratio of transmitted flux to reflected flux. This approximation is written

$$\Delta T / \Delta R = T / R \quad (6.40)$$

where T and R are the unperturbed transmittance and reflectance. The change in reflectance due to surface roughness is then

$$\Delta R = (I_{>} + I_{<})_{ns} / (1 + T / R) \quad (6.41)$$

The ratio T / R is calculated in Appendix I.

6.2 One dimensional grating.

The power scattered by a periodic, sinusoidal grating ingrained upon the surface of the antiferromagnet is now calculated. Here the surface profile function takes the form

$$\zeta(x) = h (\cos(sx) - 1) / 2 \quad (6.42)$$

The grating depth is given by h , which is assumed small in comparison to the incident wave's wavelength. The spatial period of the grating is $1/s$. Note that the expression for the grating is such that $\zeta(x) < 0$, as required by the evaluation of the integral in equation (6.16) and (6.17).

Transforming, as per equation (6.12), the profile function becomes

$$\zeta(k) = h \left[\delta(k-s) + \delta(k+s) - \frac{1}{2} \delta(k) \right] \quad (6.43)$$

When substituted into the expressions for \vec{S} given by (6.28), the delta functions pick out the Fourier coefficients of the scattered power flow corresponding to $k_{x0}+s$, $k_{x0}-s$, and k_{x0} from the integrals over k_x and k_x' . These terms correspond to the zero order reflected beam plus the first order diffracted beams. When evaluating the expressions for \vec{p} , however, one encounters terms proportional to $\exp(2isx)$ and $\exp(isx)$ which represent interference terms between the three beams. If a large portion of the grating is illuminated, the scattered beams are averaged over a large range of x . In this spatial average, the interference terms vanish.

Substituting the profile function (6.43) into the real parts of equations (6.28) the following expressions for the power flow into the material are obtained:

$$\begin{aligned} \langle \vec{S}_y \rangle = & \frac{-h^2}{64\pi^5 \omega_0^2} \left\{ \left(\epsilon_y^{-1} \vec{k}_y(k_{x0}+s) \times \vec{\lambda}_y(k_{x0}+s) \right) \times \vec{\lambda}_y^*(k_{x0}+s) e^{-2y \text{Im}(k_{y0}^*)} \right. \\ & + \left(\epsilon_y^{-1} \vec{k}_y(k_{x0}-s) \times \vec{\lambda}_y(k_{x0}-s) \right) \times \vec{\lambda}_y^*(k_{x0}-s) e^{-2y \text{Im}(k_{y0}^*)} \\ & \left. + \frac{1}{4} \left(\epsilon_y^{-1} \vec{k}_y(k_{x0}) \times \vec{\lambda}_y(k_{x0}) \right) \times \vec{\lambda}_y^*(k_{x0}) e^{-2y \text{Im}(k_{y0}^*)} \right\} \quad (6.44) \end{aligned}$$

Here we've introduced the abbreviations $k_{y>}^{\pm} = k_{y>}(k_{xo} \pm s)$ and $k_{y>}^0 = k_{y>}(k_{xo})$. The functional form for $k_{y>}$ is given explicitly by equation (6.26).

The brackets around \vec{S} indicate that the expression is averaged over x with the assumption that L_x is much larger than l/s . Outside the material, the Poynting vector is:

$$\begin{aligned} \langle \vec{S}_z \rangle = & \frac{-ch^2}{64\pi^s \omega_o^2} \left\{ \left(\vec{k}_z(k_{xo}+s) \times \vec{\lambda}_z(k_{xo}+s) \right) \times \vec{\lambda}_z^*(k_{xo}+s) e^{2y|\text{Im}(k_{y>}^+)|} \right. \\ & + \left(\vec{k}_z(k_{xo}-s) \times \vec{\lambda}_z(k_{xo}-s) \right) \times \vec{\lambda}_z^*(k_{xo}-s) e^{2y|\text{Im}(k_{y>}^-)|} \\ & \left. + \frac{1}{4} \left(\vec{k}_z(k_{xo}) \times \vec{\lambda}_z(k_{xo}) \right) \times \vec{\lambda}_z^*(k_{xo}) e^{2y|\text{Im}(k_{y>}^0)|} \right\} \end{aligned} \quad (6.45)$$

The k_x dependance of $k_{y<}$ is given by equation (6.28) and is abbreviated in the same manner as in (6.44).

Several comments are in order. First, these expressions represent power flow due to sources originating in the surface grating. These sources are driven by the incident fields. When the surface roughness is "turned off", the scattered fields must vanish. This obviously holds for the scattered power flows of (6.44) and (6.45).

When the surface roughness is "turned on", some of the incident energy is scattered out of the specularly reflected beam. This is clearly represented by the first two terms in (6.44) and (6.45), which describe scattering into the first order diffracted beams. Inclusion of higher order perturbation terms in the wave equation (6.7) for the scattered fields would result in the appearance of higher order diffracted beams in the power flow expressions.

The surface roughness does not necessarily scatter the incident energy in non-specular directions, however. The third terms in equation (6.44) and (6.45) represent

scattering into the zeroth order diffracted beam. These describe power flow scattered in the direction of the unperturbed specular beams.

For certain values of $k_{xo} \pm s$, $k_{y<}$ will be pure imaginary and will thus describe fields exponentially decaying away from the surface. These fields correspond to evanescent waves travelling along the surface perpendicular to the easy axis. Whether $k_{y<}$ is real or imaginary depends on the magnitude of k_{xo} (which depends on the angle of incident wave with respect to the surface) and the magnitude of the grating's period l/s .

When the grating induces evanescent waves travelling along the surface, there exist frequencies where surface polariton modes can be excited. In this way the grating allows for coupling between the incident electromagnetic wave and surface antiferromagnetic polaritons. To illustrate this, in figure 6.2 the dispersion curves for bulk and surface polaritons are reproduced. In this example and the rest that follow, the material is MnF_2 , a uniaxial antiferromagnet with the parameters $H_e = 7.87 \text{KGauss}$, $H_a = 550 \text{KGauss}$, $M_s = .6 \text{KGauss}$ and $\epsilon_z = 5.5$. The antiferromagnetic resonance frequency, Ω , is given by $\Omega = \gamma(2H_a H_e + H_a^2)^{1/2}$, where γ is the gyromagnetic ratio. Unless otherwise stated, damping is always present with the value $1/\Omega\tau = .0001$. The unitless frequency and wavevector are given by ω/Ω and kc/Ω , respectively. Also, the approximation discussed in chapter 5 is used to allow the Green's functions to represent leaky wave excitations.

These curves in figure 6.2 are for zero applied field. The shaded areas represent bulk polariton modes and the dotted lines rising out of the lower bulk band are surface polariton modes. The solid lines inside the bulk bands and above the surface modes are the leaky modes described in Chapter 4. The straight, nearly vertical line is the light line where $\omega_0 = k$. The two straight dashed lines correspond to the grating induced wavenumbers $k_{xo} \pm s$ for the case $k_{xo}c/\Omega = .7$ and $s = .5c/\Omega$ (In MnF_2 , with an antiferromagnetic resonance frequency $\Omega = 268 \text{GHz}$, this s is approximately $.5 \text{mm}$). For these values, $k_{y<}$ is pure imaginary for the $+s$ case and the corresponding scattered fields decay exponentially away from the surface.

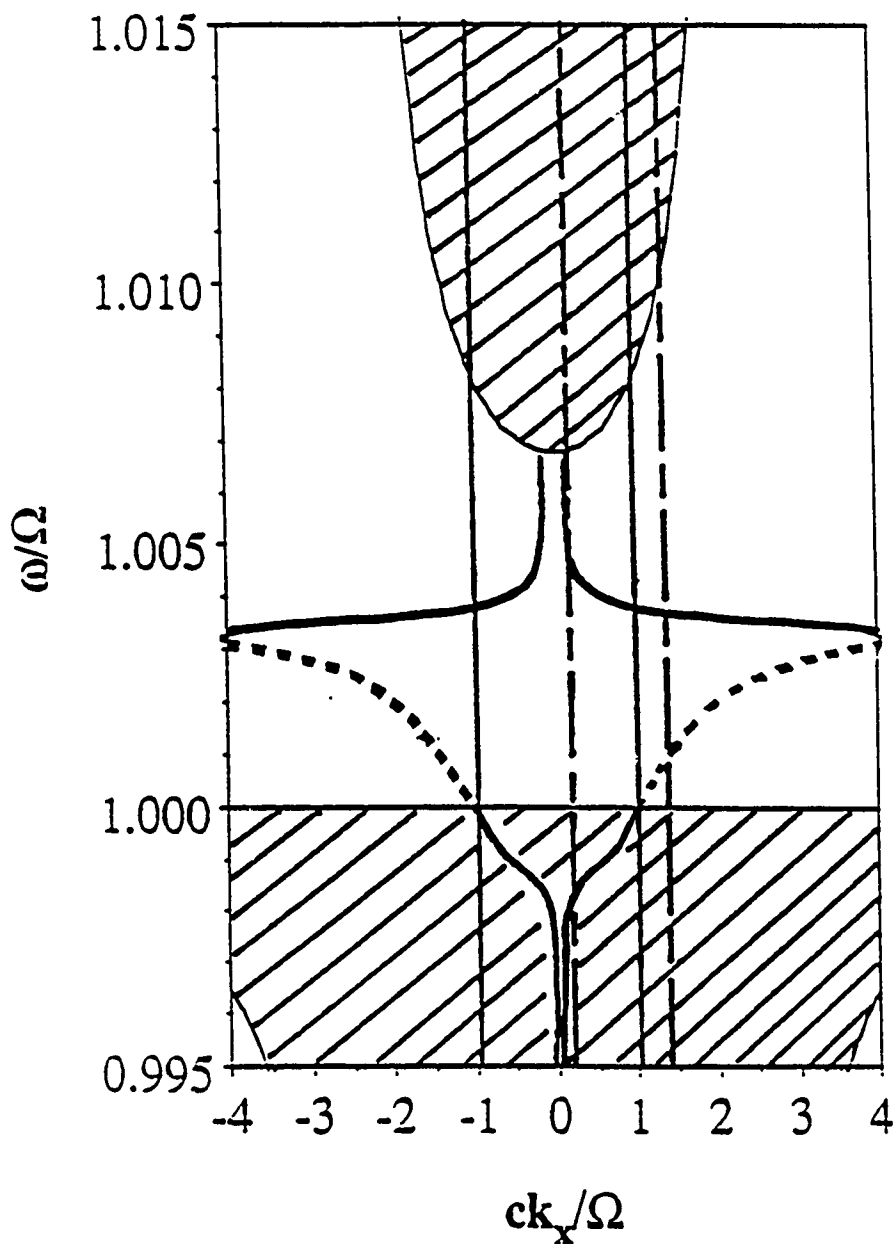


Figure 6.2. Dispersion curve for antiferromagnetic polaritons, in MnF_2 , with no applied field. The shaded areas are bulk bands and the dotted lines are surface modes. The solid lines are leaky modes. The straight solid lines are light lines, where $\omega_0 = k_x$. The straight dashed lines are the grating induced lines, $q = k_x \pm s$, for $k_x c/\Omega = .7$ and $s = .5c/\Omega$.

The light line does not intersect any of the surface polariton modes, but for this choice of s , the grating induced line of $k_{x0} + s$ intersects a surface mode. The presence of the grating thus allows an incident electromagnetic wave to couple with polariton modes inaccessible when the surface is ideally smooth. Note that higher order terms in the power flow expressions would result in more grating induced lines being drawn, and hence couplings to portions of the surface polariton modes at even shorter wavelengths.

The sum of the $+s$ and the $-s$ terms for the parallel power flows are plotted in figure 6.3 for $H_0=0$, $s=.5c/\Omega$ and $\theta_0=45^\circ$. Plotted are $\langle S_x^< \rangle / ch^2$ just outside the material in vacuum at $y=0^-$ and $\langle S_x^> \rangle / ch^2$ just inside the material at $y=0^+$. The large peak at $\omega = 1.001\Omega$ is approximately where the grating line intersects the surface mode shown in the dispersion curves of figure 6. Note that the energy flow outside the material is larger and oppositely directed to the energy flow inside the material. Inside the material, the susceptibilities are negative for the surface modes. In consequence, the Poynting vector is also negative and oppositely directed to the vacuum Poynting vector. This observation has also been made for surface plasmon-polaritons and surface magnetoelastic polaritons.

The smaller peak in figure 6.3 is from a surface resonance that lies *within* the bulk band. This peak corresponds to a location on the dispersion curve of figure 6.2 at $\omega = .998\Omega$ on the $-s$ grating line. Note that this places the resonance on the left side of the light line, in a region where true surface polaritons cannot exist. These resonances are due to the Brewster-like leaky modes described in Chapter 4. A similar strong excitation of a leaky mode due to the coupling provided by a grating has been discussed for the case of elastic waves by Glass and Maradudin.

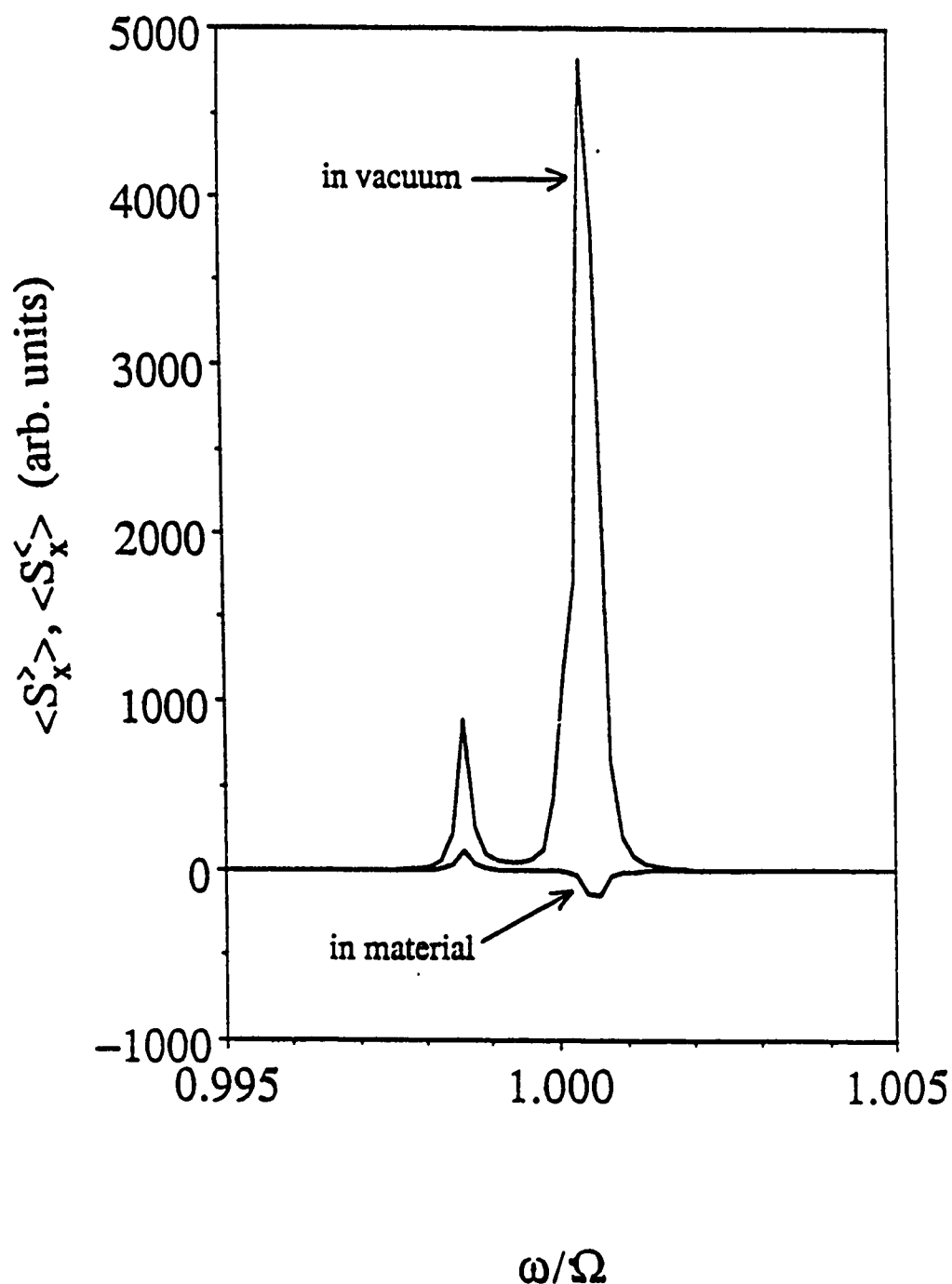


Figure 6.3. Parallel power flows inside and outside the material as functions of frequency on a periodic grating with $s=.5c/\Omega$. The quantities shown are proportional to $\langle S_x^{\rangle}$ and $\langle S_x^{\lessgtr}>$. There is no applied field, $\theta_0=45^\circ$, and damping is $1/\Omega\tau=.0001$. The larger, higher frequency peak is due to a surface polariton. The other peak results from a surface resonance that lies within the bulk polariton band.

The leaky waves depend on material damping for their existence and so it is interesting to study them for larger damping parameters. In figure 6.4 the parallel power flows are plotted for the same values of s and θ_0 as in figure 6.3 but with a larger damping constant of .0008. The quantity actually plotted is $\langle S_x \rangle / ch^2$. The larger peak is still the surface polariton but there is a distinct broadening of both the resonance and the polariton. In both figures 6.3 and 6.4 also note the power flows inside the material are opposite those outside the material for both surface excitations. While the surface polariton carries most of its energy outside the material, the resonance carries most of its energy inside the material.

Having examined the possible energy flows inside and outside the material, the fraction of energy scattered out of an incident wave into these radiative and evanescent states can be calculated. For the periodic grating the power ratio equations (6.33, 6.35, 6.36, and 6.37) become

$$I_{>}^r = \frac{h^2}{16H^{(i)2} \pi^4 \omega_o^2 \epsilon_2 \cos(\theta_0)} (q\lambda_{y>}(q) - k_{y>}(q)\lambda_{x>}(q))\lambda_{x>}^*(q) \quad (6.46)$$

$$I_{<}^r = \frac{h^2}{16H^{(i)2} \pi^4 \omega_o^2 \cos(\theta_0)} (q\lambda_{y<}(q) - k_{y<}(q)\lambda_{x<}(q))\lambda_{x<}^*(q) \quad (6.47)$$

$$I_{>}^e = \frac{h^2}{32L_z |\text{Im}(k_{y>}(q))| H^{(i)2} \pi^4 \omega_o^2 \epsilon_2 \cos(\theta_0)} (k_{y>}(q)\lambda_{x>}(q) - q\lambda_{y>}(q))\lambda_{y>}^*(q) \quad (6.48)$$

$$I_{<}^e = \frac{h^2}{32L_z |\text{Im}(k_{y<}(q))| H^{(i)2} \pi^4 \omega_o^2 \cos(\theta_0)} (k_{y<}(q)\lambda_{x<}(q) - q\lambda_{y<}(q))\lambda_{y<}^*(q) \quad (6.49)$$

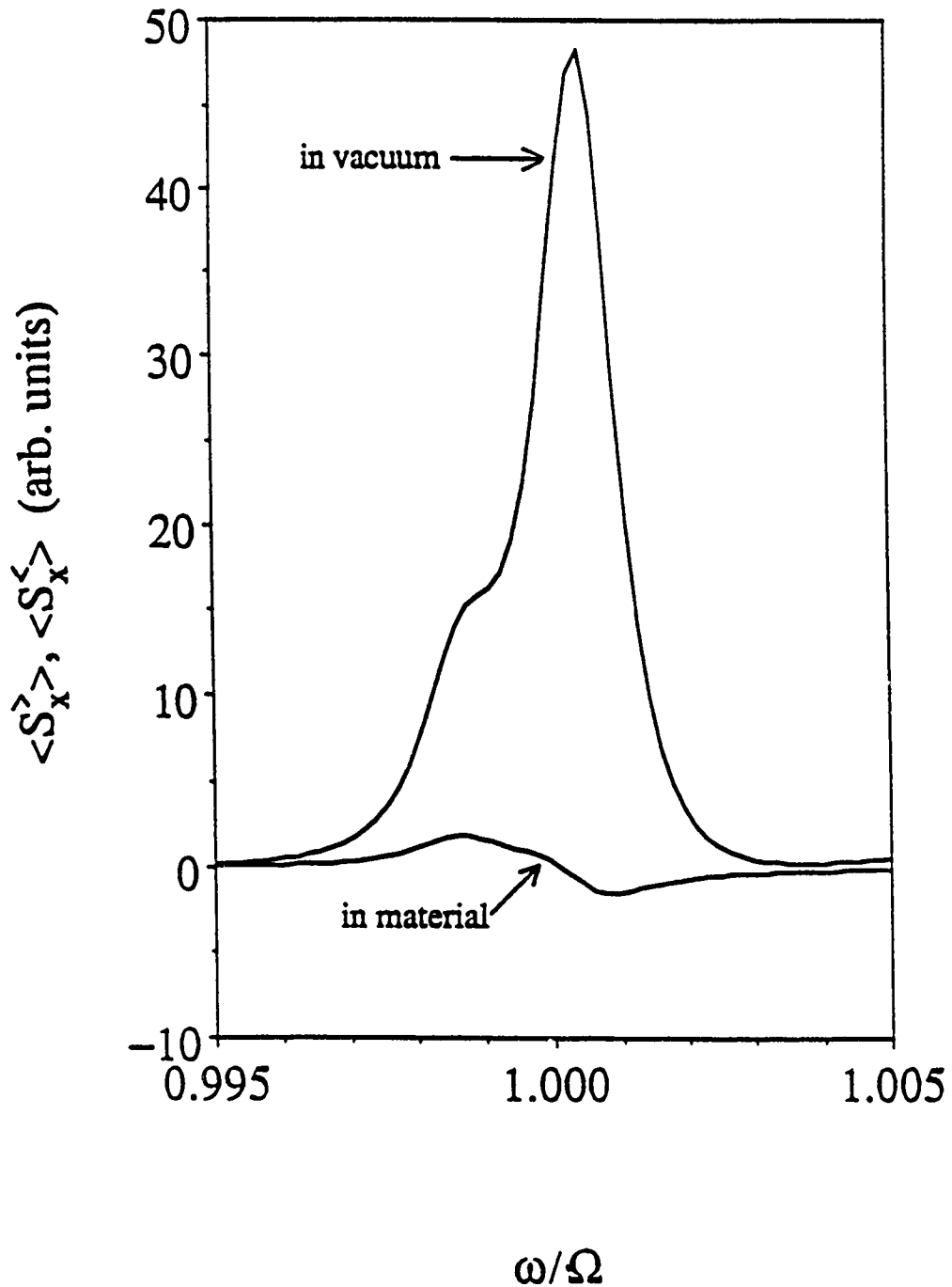


Figure 6.4. Parallel power flows inside and outside the material as functions of frequency on a periodic grating with $s = .5c/\Omega$. The quantities shown are proportional to $\langle S_x^{\rightarrow} \rangle$ and $\langle S_x^{\leftarrow} \rangle$. There is no applied field, $\theta_0 = 45^\circ$, and damping is now $1/\Omega\tau = .0008$. The peaks are much broader now than in the .0001 damping case of figure 6.3.

Here q is understood as either k_{x0} , $k_{x0}+s$ or $k_{x0}-s$. As discussed earlier, the magnitude of q determines whether the power flow is radiative or evanescent. Thus $I_{>}^r$ represents radiative scattering in the material when $|q| < |\vec{k}_{>}|$ and $I_{>}^e$ represents evanescent scattering $|q| > |\vec{k}_{>}|$. Similarly, $I_{<}^r$ and $I_{<}^e$ represent radiative and evanescent scattering in vacuum when $|q| < |\vec{k}_{<}|$ and $|q| > |\vec{k}_{<}|$.

Before proceeding, recall that in an applied field the surface modes become highly nonreciprocal with respect to propagation direction (i.e., $\omega(+k_x) \neq \omega(-k_x)$). For future reference, the dispersion curves for the bulk and surface polaritons in an applied field of .3kG are presented in figure 6.5 for $1/\Omega\tau = .0001$. The shaded areas are again bulk polariton bands and the dashed lines are surface polariton modes. The solid lines are surface leaky modes.

Now it is possible, for certain choices of grating period s , to couple an incident wave whose parallel wavevector component is in the $+x$ direction with surface modes whose wavevectors are in the $-x$ direction. For example, suppose a wave of frequency ω incident on a grating with period $1/s$ at some angle θ_0 couples to a surface mode travelling in the $-x$ direction with wavevector $k_{x0}-s$. It is also possible for an incident wave at another frequency ω' to couple with a surface mode travelling in the $+x$ direction with wavevector $k_{x0}'+s$ for the same angle of incidence. Thus for fixed s and θ_0 , one can couple to both the $+k_x$ and the $-k_x$ polariton branches by scanning the frequency of the incident wave. Placing the material in applied field will increase the difference in frequencies by exploiting the nonreciprocity of the surface modes. An example is given in figure 6.6 where the evanescent power flows are plotted as functions of frequency for $s=3c/\Omega$, and $\theta_0=-45^\circ$. An external field of .3kG is applied. The quantities actually plotted are $I_{>}^e L_z / \hbar^2$.

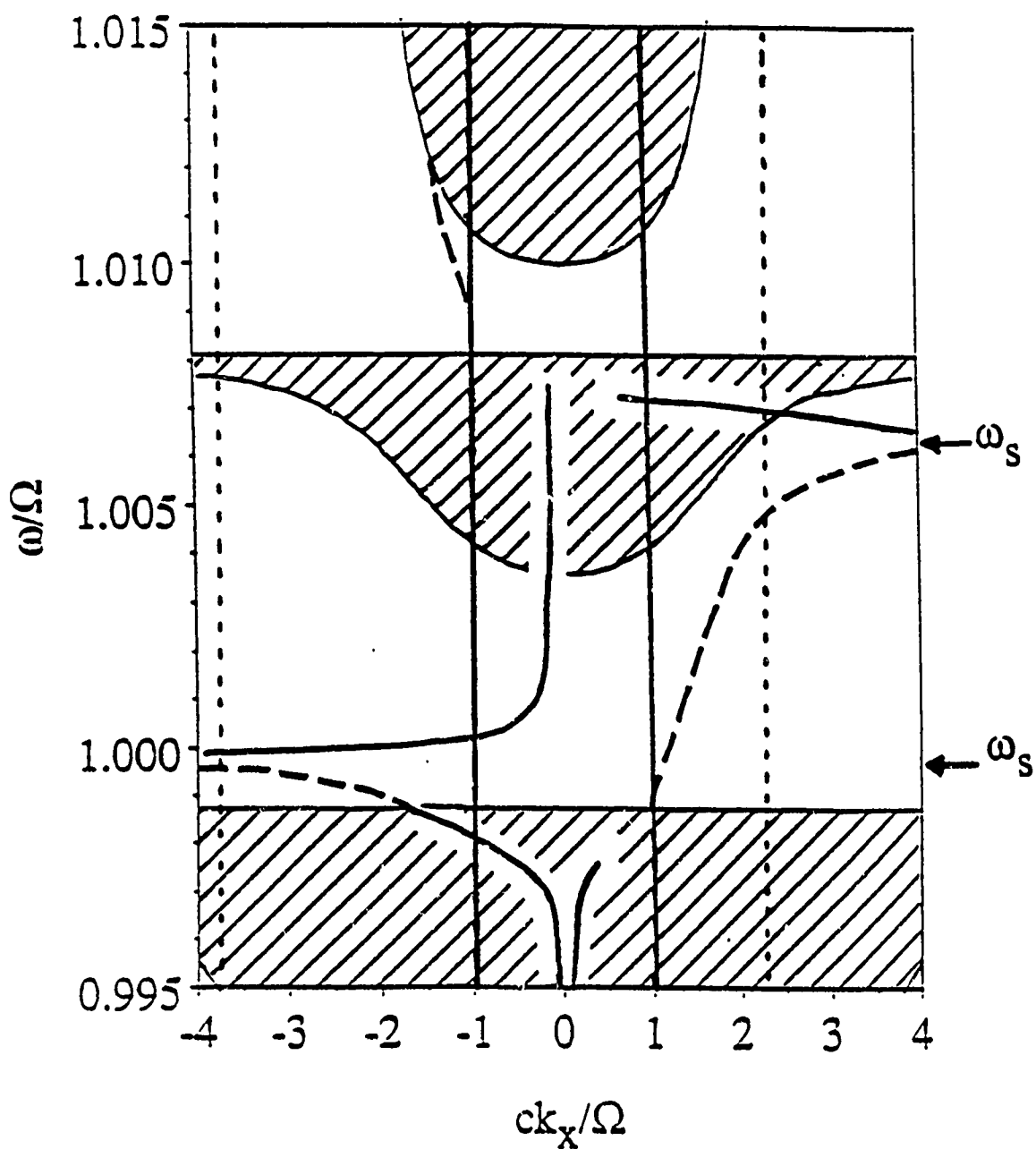


Figure 6.5. Dispersion curve for antiferromagnetic polaritons, in MnF_2 , with an applied field of .3kG. The shaded areas are bulk bands and the dotted lines are surface modes. The solid lines are leaky surface modes. The straight solid lines are the light lines, where $\omega_0 = k_x$. The straight dashed lines are the grating induced lines, $q = k_x \pm s$, for $k_x c/\Omega = -0.7$ and $s = 3c/\Omega$. Note the strong nonreciprocity of the surface modes with respect to propagation direction.

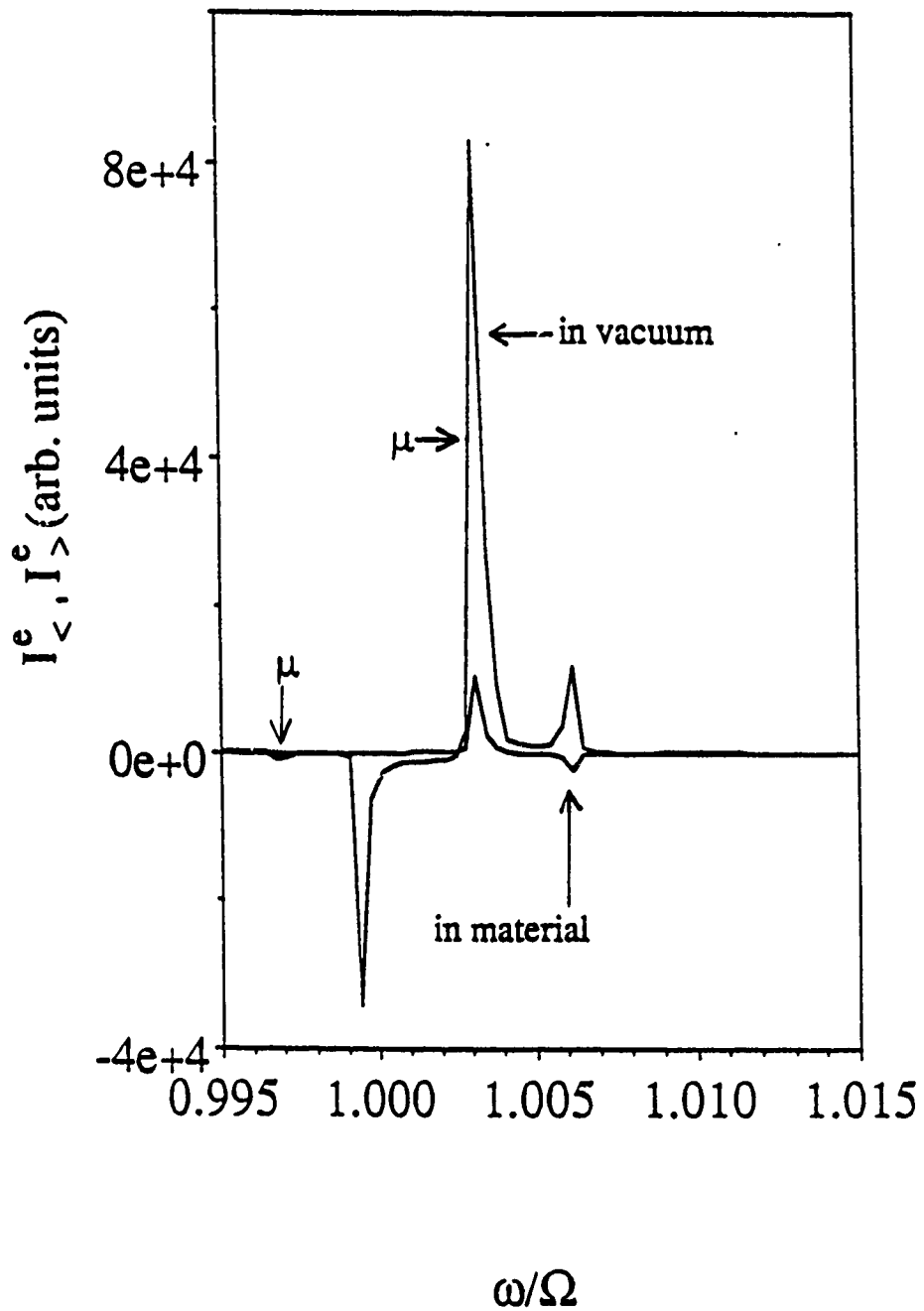


Figure 6.6. Evanescent power flows inside and outside the material as functions of frequency on a periodic grating with $s=3c/\Omega$. The quantities shown are proportional to $I_>^e$ and $I_<^e$. The applied field is .3kG, $\theta_0=-45^\circ$, and damping is $1/\Omega\tau = .0001$. For this s , we now pick up both the $+k_x$ surface mode and the $-k_x$ surface mode. The peaks labelled μ occur at frequencies where the susceptibilities become very large.

To identify the peaks in this figure, refer to the dispersion curves of figure 5 where there are lines representing the grating induced wavevectors for $s = 3c/\Omega$ and $k_{xo}c/\Omega = -.7$. The peak nearest $\omega/\Omega = 1$ in figure 6 corresponds to where the $-s$ grating line crosses the $-k_x$ surface polariton mode in figure 6.6. The peak at $\omega = 1.005\Omega$ in figure 6.6 corresponds to where the $+s$ grating line crosses the $+k_x$ surface polariton mode.

The power flow for both modes is largest in vacuum and in the directions expected for the different modes: I_z^e is positive for the $+k_x$ mode and negative for the $-k_x$ mode. Also, note that the power flow in the material is no longer in a direction opposite that of the vacuum flow. This does not contradict the zero field case because although μ_1 is still negative, the applied field also influences the direction of the material power flow through the off diagonal element of the susceptibility tensor, μ_2 .

The two peaks labelled " μ " are at the antiferromagnetic resonance frequencies. At these frequencies μ_1 and μ_2 become very large and change sign. Since the driving fields of the perturbation expansion are proportional to μ_1 and μ_2 , the theory breaks down at these frequencies. Thus the magnitude of the energy flow in the evanescent (and radiative) fields at these frequencies may be exaggerated with respect to the magnitude of the surface polariton peaks.

In figure 6.7 the reflectance of the smooth surface, R , and the reflectance of the rough surface, $R - \Delta R$, is plotted vs θ_o for two frequencies. There is no applied field and $s = 1c/\Omega$. The grating height is $h = .0002c/\Omega$ (for MnF_2 , this height is approximately $.2\mu\text{m}$) and the width of the illuminating beam, L_x , is set equal to Ω/c for convenience. The presence of the grating introduces dips in to the reflectance which are not seen for the smooth surface. These dips are, as usual, due to the coupling to surface polaritons, in this case the "true" surface polaritons, not the surface resonances. The dips are reciprocal in angle and are much further apart at the higher frequency than at the lower. This is due to the positive group velocity of the modes, as seen in figure 6.2, where the higher frequency modes exist at larger wavevectors.

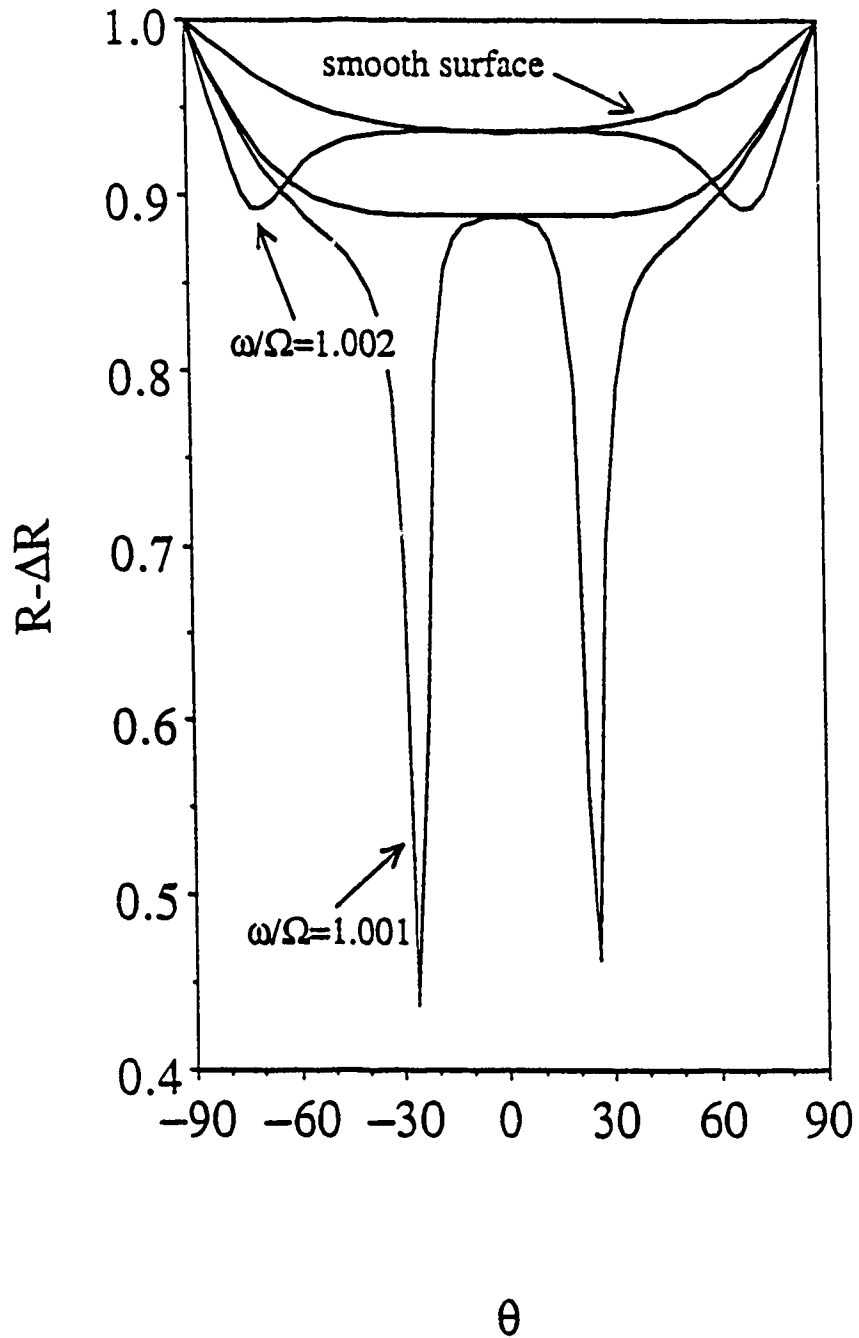


Figure 6.7. Reflectivity as a function of θ_0 for a smooth surface (dashed lines) and a surface with a grating of period $s=1c/\Omega$ (solid lines). There is no applied field and damping is $1/\Omega\tau=.0001$. The two frequencies couple with surface polaritons with different group velocities. The higher frequency polariton has a smaller group velocity and consequently has a greater density of k_x states. Thus it produces a broader and deeper reflectivity dip than the lower frequency mode.

An interesting point is the increased width of the dips at the higher frequency. This is due to the group velocity, which decreases with increasing frequency, thus creating a higher density of states at the higher frequency. Since damping allows the incident wave to excite surface polaritons over a small range of frequency, a greater density of states allows the incident wave to excite a larger number of polariton states.

6.3 Randomly rough surface.

The case of random roughness is somewhat simpler algebraically than the one dimensional grating. First, the average of S over the possible height profiles $\zeta(x)$ is calculated. A gaussian form for the correlation function is assumed:²⁶

$$\langle \zeta(x) \zeta^*(x+x') \rangle = h^2 e^{-x'^2/\sigma^2} \quad (6.50)$$

Here h corresponds to the mean square height $\langle \zeta^2(x) \rangle$ and σ is the correlation length for the distribution of profiles. The fourier transform is simply

$$\langle \zeta(k) \zeta^*(k) \rangle = \frac{1}{4\pi} h^2 \sigma^2 e^{-k^2 \sigma^2/4} \delta(k-k') \quad (6.51)$$

Using the profile correlation function of (6.51), the power ratio equations (6.33, 6.35, 6.36, and 6.37) are given by

$$I_{>}^r = \frac{h^2 \sigma^2}{64 H^{(i)2} \pi \omega_0^2 \epsilon_2 \cos(\theta_0)} \int_{-|\vec{k}_{>}|}^{|\vec{k}_{>}|} dk_x (k_x \lambda_{y>}(k_x) - k_{y>} \lambda_{x>}(k_x)) \lambda_{x>}^* e^{-(k_x - k_{x0})^2 \sigma^2/4} \quad (6.52)$$

and

$$I_{<}^r = \frac{\hbar^2 \sigma^2}{64 H^{(i)2} \pi^3 \omega_0^2 \cos(\theta_0)} \int_{|\vec{k}_{<}|}^{|\vec{k}_{>}|} dk_x (k_x \lambda_{y<}(k_x) - k_{y<} \lambda_{x<}(k_x)) \lambda_{x<}^*(k_x) e^{-(k_x - k_{x0})^2 \sigma^2 / 4} \quad (6.53)$$

for the radiative fields, and

$$I_{>}^e = \frac{\hbar^2 \sigma^2}{128 L_x H^{(i)2} \pi^4 \omega_0^2 \epsilon_2 \cos(\theta_0)} \int_{|k_x| > |\vec{k}_{>}|} dk_x (k_{y>} \lambda_{x>}(k_x) - k_x \lambda_{y>}(k_x)) \lambda_{y>}^*(k_x) \frac{e^{-(k_x - k_{x0})^2 \sigma^2 / 4}}{|\text{Im}(k_{y>})|} \quad (6.54)$$

$$I_{<}^e = \frac{\hbar^2 \sigma^2}{128 L_x H^{(i)2} \pi^4 \omega_0^2 \cos(\theta_0)} \int_{|k_x| > |\vec{k}_{<}|} dk_x (k_{y<} \lambda_{x<}(k_x) - k_x \lambda_{y<}(k_x)) \lambda_{y<}^*(k_x) \frac{e^{-(k_x - k_{x0})^2 \sigma^2 / 4}}{|\text{Im}(k_{y<})|} \quad (6.55)$$

for the evanescent fields.

The change in reflectivity is calculated by (6.41) as for the periodic grating. The nonspecular portions of $I_{>}^r$ and $I_{<}^r$ are calculated by excluding from the integration a small region about k_{x0} . This exclusion has a physical interpretation when the Fourier amplitudes at each k_x are related to a scattering angle, θ_s , via the geometrical relation

$$k_x = |\vec{k}_{>}| \sin \theta_s \quad (6.56)$$

In a reflection experiment, the excluded region would be the angular width of a detector.

This width is taken as one degree and the k_x values in the range $|\vec{k}_{>}| \sin(\theta_0 + 1/2)$ to $|\vec{k}_{>}| \sin(\theta_0 - 1/2)$ are excluded from the integrations in (6.52-6.55).

With random roughness, there is an infinite collection of grating periods which produce an effective "width" in wavelength analogous to the damping width in frequency. This width is measured by the correlation length σ . Smaller values of σ lead to a greater smearing in wavelength while larger values indicate a smoother surface and a narrower wavelength range.

The dependence of the scattering on σ is shown in figures 6.8 and 6.9. In both figures I_{ζ}^{θ} is plotted as a function of frequency for $\theta_o = -45$. There is an applied field of .3kG. In figure 8, $\sigma = .1c/\Omega$ (for MnF_2 , this correlation length is approximately .1mm). The largest peak represents coupling with the $-k_x$ surface mode of figure 6.5. The smaller peak at the higher frequency represents coupling with the $+k_x$ surface mode. In both cases the largest coupling occurs where the density of states is the greatest. In figure 6.5 this occurs where the surface branches flatten out and the group velocity approaches zero. This occurs at rather large wavevectors and so requires a fairly rough surface to allow coupling with an incident wave. The " μ " peaks discussed earlier are present, but dwarfed in comparison to the surface mode peaks.

In figure 6.9, $\sigma = 1c/\Omega$ (a correlation length of about 1mm for MnF_2) and a distinctly different profile for the scattered energy results. First, the magnitude of the surface mode peaks is considerably reduced in comparison to the " μ " peaks which are now clearly visible at $\omega_o/\Omega = .997$ and $\omega_o/\Omega = 1.003$. The surface mode peaks are also at lower frequencies than the rougher surface case. The shift in frequency and the reduced amplitudes of these surface mode peaks are both due to lack of coupling to the higher (and denser) frequency portions of the surface mode dispersion curves. The strongest coupling for this smoother surface now occurs nearer to the light line.

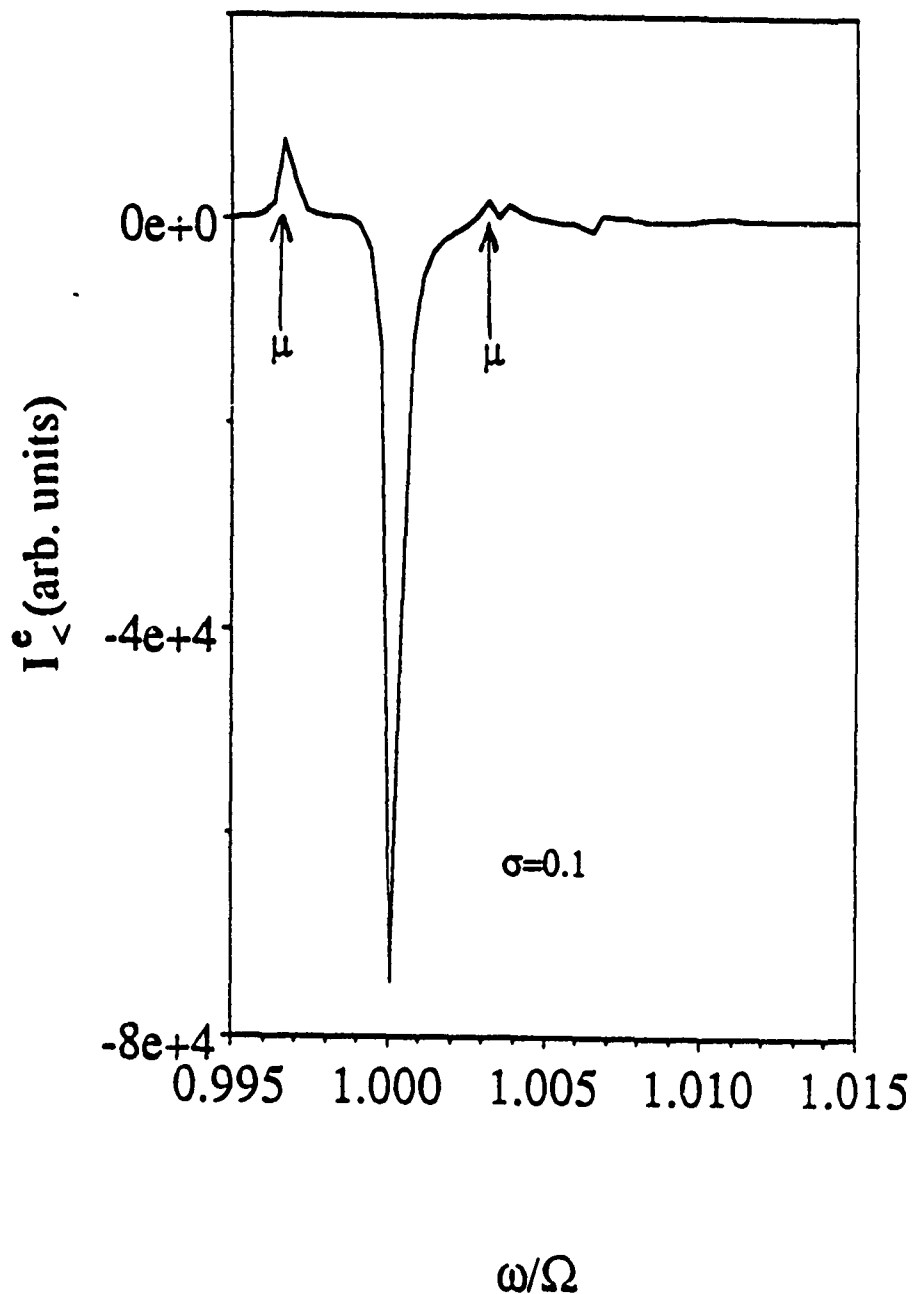


Figure 6.8. Evanescent power flows outside the material as functions of frequency on a rough surface with $\sigma=1c/\Omega$. The quantity shown is proportional to I_z^e . The applied field is .3kG, $\theta_0=-45^\circ$, and damping is $1/\Omega\tau=.0001$. For this σ , we pick up both the $+k_x$ surface mode and the $-k_x$ surface mode near their limiting frequencies where the density of states is largest.

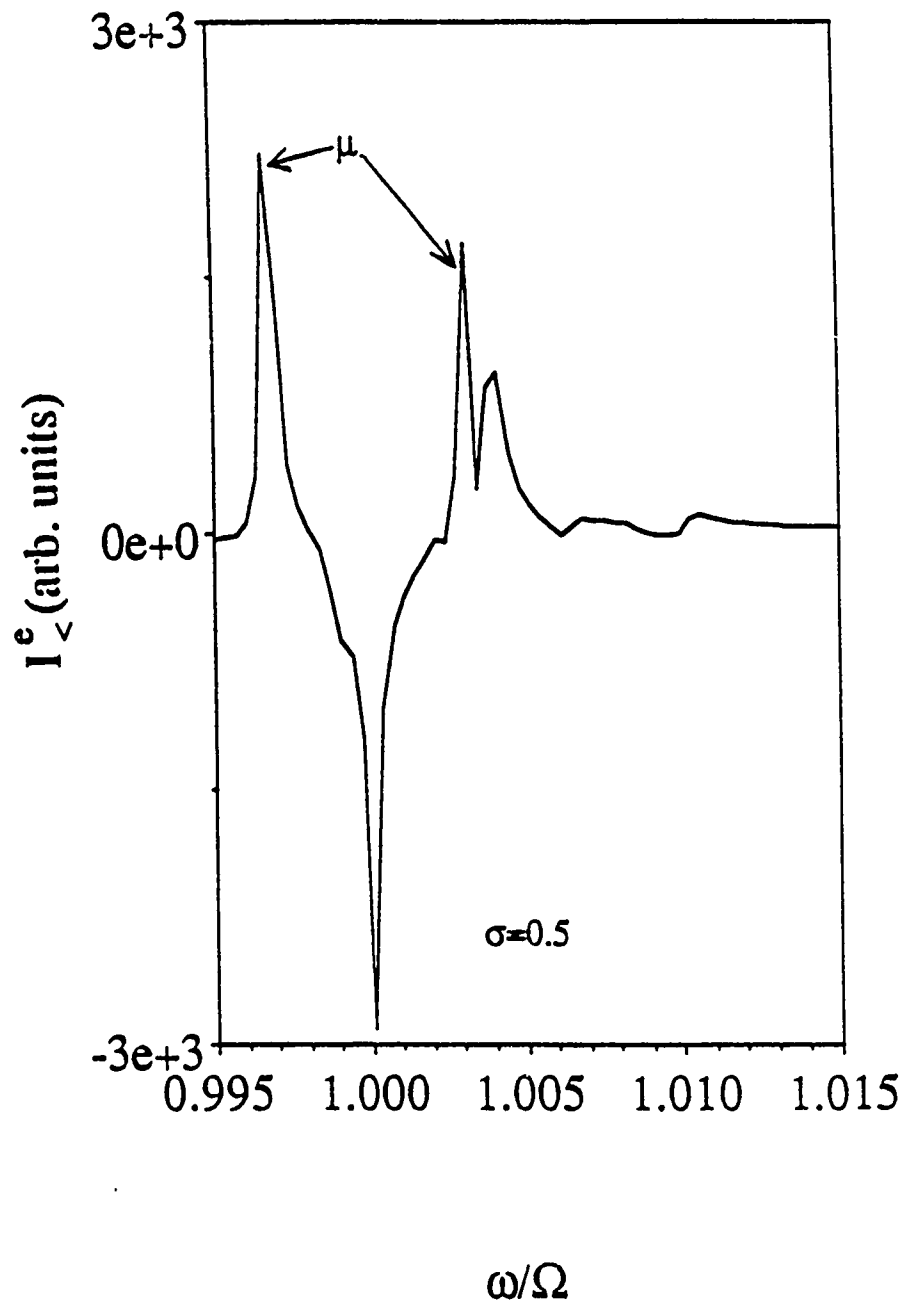


Figure 6.9. Evanescent power flows outside the material as functions of frequency on a rough surface with $\sigma = .5c/\Omega$. The quantity shown is proportional to I_z^e . The applied field is .3kG, $\theta_0 = -45^\circ$, and damping is $1/\Omega\tau = .0001$. For this σ , we pick up contributions from the surface polaritons nearer the light line. These polaritons have relatively large group velocities and small state densities, hence they are not as strong scatters as the shorter wavelength polaritons. Note the prominence of the " μ " peaks compared to the surface polariton peaks.

In figure 6.10 the effect of σ on a reflection measurement is explored. Here the reflectance of the smooth surface, R , is shown together with the reflectance of the rough surface, $R-\Delta R$, as functions of θ_0 in an applied field of .3kG. The frequency is $.9989\Omega$ and so the incident wave can couple with both the $+k_x$ and the $-k_x$ surface mode branches of figure 5. The rms height is $h=.003c/\Omega$ (a height of $3\mu\text{m}$ for MnF_2) and the cases $\sigma = .005c/\Omega$ and $\sigma = .01c/\Omega$ are presented. Note that with this small damping value, the reflectance from the smooth surface is fairly reciprocal with respect to θ_0 . The rough surface, however, induces nonreciprocity by coupling the incident wave more strongly with the surface polariton modes.

The greatest losses occur for couplings with the $-k_x$ branch. This is because the $-k_x$ branch is flatter than the $+$ branch, and thus has a greater density of states nearer the light line than the $+k_x$ branch. Not only has the change in reflectivity increased with the slightly rougher surface, as expected by the σ^2 dependance of ΔR , but the $+$ and $-$ peaks are also more pronounced.

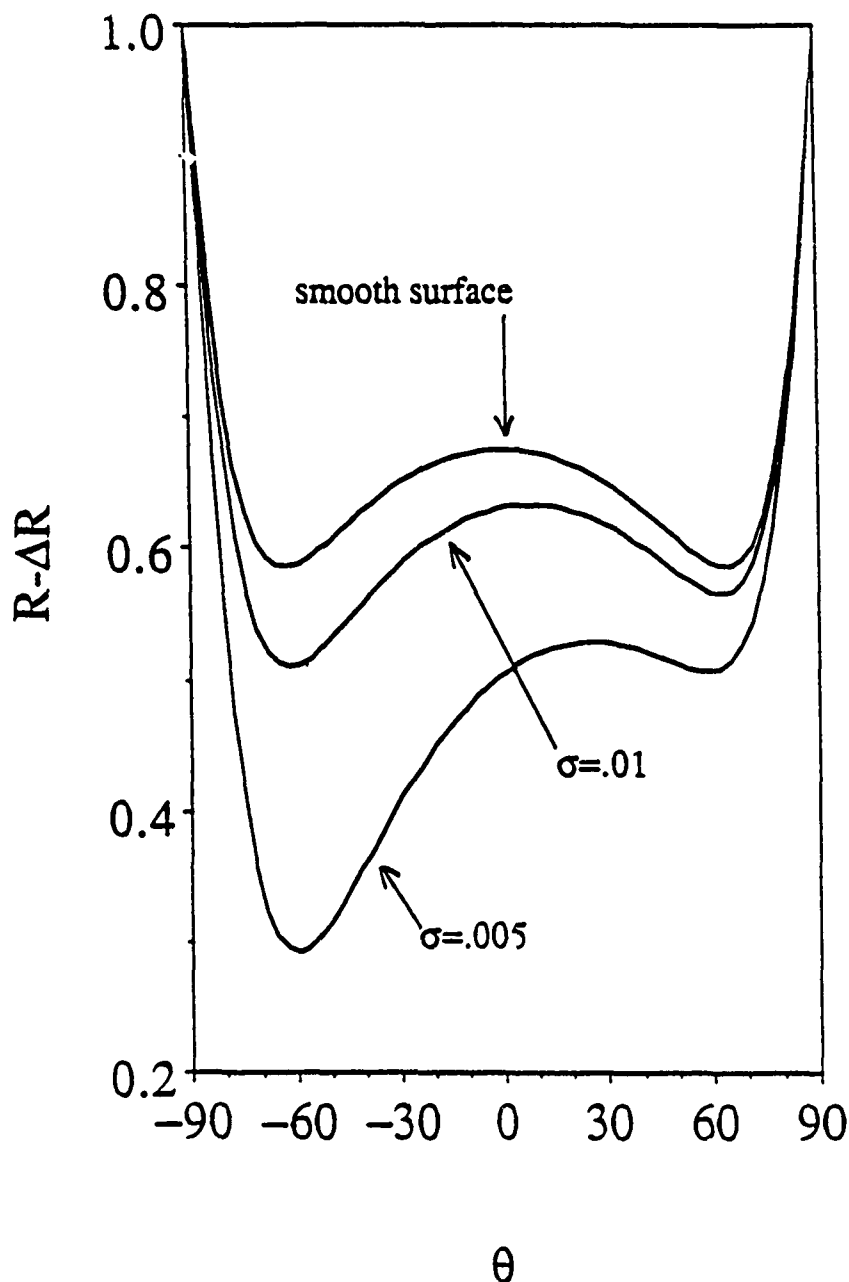


Figure 6.10. Reflectivity as a function of θ_0 for a smooth surface, a rough surface with $\sigma=1c/\Omega$, and a rough surface with $\sigma=.1c/\Omega$. There is an applied field of .3kG and damping is $1/\Omega\tau=.0001$. The rms height is $h=.003c/\Omega$. The frequency of the incident wave is .9989kG which allows coupling to both the $+k_x$ and the $-k_x$ surface polariton. With the applied field, these modes have different group velocities. The largest dip is due to the $+k_x$ surface polariton which has the greatest density of states nearest the light line. As roughness is increased, the nonreciprocal reflection losses become more pronounced.

CHAPTER 7

CONCLUSIONS

In this work, electromagnetic excitations on semi-infinite antiferromagnets have been examined within the framework of a phenomenological theory based on the macroscopic Maxwell equations. The problem was approached from two points of view; (1) a description of the possible excitations of the electromagnetic field created by an antiferromagnet through dispersion relations obtained by solving the wave equation, and (2) a representation depicting the propagation of electromagnetic waves on the antiferromagnet obtained by calculating the surface response functions appropriate to the semi-infinite geometry.

Throughout the discussion, the focus has been primarily on topics of interest to possible experimental and technological application. For this reason the coupling between photons and polaritons has been considered in detail and mathematical models have been constructed that describe possible experimental situations. Four investigations were carried out. The first concerned the possibility of nonreciprocal reflection from antiferromagnets when absorption mechanisms were present. The second investigation revealed that when damping is present in the antiferromagnet, new surface excitations may exist that are quite distinct from antiferromagnetic surface polaritons. The third investigation showed that incident light could excite surface excitations by scattering off a sinusoidal grating ruled on the surface of the antiferromagnet. The final investigation showed how nonreciprocal reflectance, originating this time in nonreciprocal absorption by surface polaritons, could be enhanced by scattering incident light from a randomly rough surface. The results of each of these four inquiries are now briefly summarized.

7.1 Bulk polaritons and nonreciprocal reflection.

A simple thermodynamic argument shows that reflection from a smooth antiferromagnet, in regions where the material is optically transparent, can be nonreciprocal with respect to incident angle even when the material is non-absorptive. By considering all the possible polarizations of the incident and reflected light, the total reflected energy summed over all polarization states is in fact reciprocal with respect to propagation angle. When the polarization states are considered individually, however, a given polarization state may have a highly nonreciprocal reflectance.

In a numerical example, the antiferromagnet MnF_2 displayed nonreciprocal reflection within circularly polarized states; i.e., the reflectance of a right circularly polarized incident wave changed as much as 20% between the incident angles $\theta=45^\circ$ and $\theta=-45^\circ$ when the antiferromagnet was in an applied field of .3kG. The nonreciprocity is strongly dependant on the frequency of the incident wave and is greatest near the antiferromagnetic resonance frequencies. In some frequency regions, the nonreciprocity will reverse with incident angle and $R(\theta)-R(-\theta)$ will change sign.

The nonreciprocity is ultimately due to a propensity of the incident wave to couple most strongly with bulk polariton modes with the same polarization. The bulk modes are right and left circularly polarized, and in an applied field bulk waves travelling in the same direction parallel to the applied field with different polarizations have different amplitudes. The polarization of the bulk modes is strongly influenced by applied magnetic fields and thus leads to nonreciprocal transmittance, and consequently, nonreciprocal reflectance of the incident wave.

7.2 Leaky surface modes on antiferromagnets.

When damping is included into the susceptibilities governing the magnetic properties of the antiferromagnet, an analysis of the resulting surface polariton dispersion

relation show surface resonances in frequency regions forbidden to "true" antiferromagnetic surface polaritons. These excitations are analogous to the leaky modes found for surface electromagnetic waves on dielectrics and are characterized as dissipative waves that "leak" energy from the surface wave into the interior of the material.

A close correspondence is found between these antiferromagnetic surface resonances and the evanescent Brewster and evanescent modes found in plasmon polariton studies. The antiferromagnetic modes have finite path lengths and can be highly dissipative with large penetration depths into the material. The penetration depth is frequency dependant and sensitive to the particular type of damping used in the description (i.e., Bloch damping versus Landau).

The surface modes with damping present are reciprocal when there is no applied field. In an applied field, their reciprocity is frequency dependant and they are nearly reciprocal at low frequencies within the lowest bulk band but become nonreciprocal at higher frequencies where they exist in the "true" surface polariton regime. Also, the direction of leaky modes' energy flow inside the material is extremely sensitive to damping and applied fields especially near the antiferromagnetic resonance frequencies.

7.3 Scattering from periodic gratings.

The electromagnetic Green's functions for a semi-infinite antiferromagnet were calculated for the special case of propagation perpendicular to the easy axis. A study of the peaks of the green's functions revealed the expected surface excitations and a correspondence between the poles of the Green's functions and the surface modes was established. The Green's functions were next applied to a perturbative treatment of scattering from two types of surface roughness: a periodic grating and a surface with random roughness in one dimension.

For certain choices of grating periods and incident angles, a periodic grating can induce coupling between incident electromagnetic waves and surface polaritons by creating

evanescent fields that travel parallel to the surface of the material. These evanescent fields are confined to the surface of the material and can have frequencies and wavevectors corresponding to surface polariton modes. On the other hand, a grating can also diffract the incident waves into radiative states that, when damping is present in the material, can couple to the leaky Brewster-like modes.

7.4 Scattering from randomly rough surfaces.

Random roughness creates a "width" in wavelength analogous to the width in frequency created by material damping. As the surface becomes rougher, the incident wave couples strongly to surface modes over a greater range of wavelengths. The surface polaritons have a greater density of states at short wavelengths, and increased roughness can lead to stronger couplings by allowing the incident wave to interact with short wavelength surface polaritons. The density of states of the surface polariton modes is also nonreciprocal with respect to propagation direction, so increased roughness in a reflection experiment can also enhance nonreciprocal reflectivity changes.

7.5 Future extensions.

The most natural extension for this work would be the experimental investigation of the phenomena predicted by these calculations. Observation of the leaky modes would be particularly interesting and the categorization of stable frequency regions for these modes, as discussed in Chapter 4, could be used to gain information about the damping mechanisms in the antiferromagnet. Observation of the leaky modes at different temperatures would be particularly interesting since different types of damping have different temperature dependences. If the leaky modes were primarily due to interactions within the spin system, as represented phenomenologically by Landau damping terms, they should be

observable at very low temperatures where magnon-phonon interactions are negligible. Conversely, if the leaky modes are due primarily to interactions involving the vibrational modes of the lattice, as described phenomenologically with Bloch damping terms, then the leaky modes should show a strong temperature dependence.

Further theoretical investigation of the leaky modes is also in order. While these excitations are not true eigenmodes of the antiferromagnetic system, the electromagnetic Green's functions can be used to approximately describe coupling to these waves. It remains to clarify and understand this approximation. Since the approximation requires exponentially increasing waves to obtain leaky mode peaks from the Green's functions, it might be interesting to calculate the exact Green's functions for a bounded geometry that would admit exponentially growing waves as physical representations. The inclusion of a ground plane in the vacuum outside the antiferromagnet would be the simplest system to consider.

There are numerous other ways to expand upon the theoretical work presented in this manuscript. For example, the Green's functions could be calculated, at least numerically, for arbitrary angles of propagation with respect to the applied field. Since the coupling of external photons to bulk polariton modes is extremely sensitive to the wavevector component of the incident wave parallel to the applied field (k_z), perhaps coupling incident light to the surface excitations is also strongly dependent on k_z . This might be particularly true for the leaky modes since they are often largely radiative in character. It would thus be interesting to examine the profiles of the Green's function peaks for the surface excitations as functions of k_z .

Finally, there remains the question of how to accurately describe the scattering of light from a rough surface. An important enlargement of the theory would be to include into the description the possibility of scattering the incident light into different polarization states. This would involve calculating the $k_z \neq 0$ Green's functions, however, since magnetic fields in the z direction are uncoupled from the other magnetic fields when $k_z = 0$.

REFERENCES

1. See the book Electromagnetic Surface Modes edited by A. D. Boardman (Wiley Interscience, New York, 1982).
2. G. N. Zhihin, M. A. Moskalova, A. A. Sigarev and V. A. Yakovelev, *Optics Communications*, 43 32 (1982).
3. See the review article by E. F. Sarmiento and D. R. Tilley in Reference 1.
4. A. Hartstein, E. Burstein, A. A. Maradudin, R. Brewer and R. F. Wallis, *J. Phys. C* 6 1266 (1973).
5. L. Remer, B. Lüthi, H. Sauer, R. Beick and R. E. Camley, *Phys. Rev. Lett.* 56 2752 (1986). There is some earlier work on bulk polaritons in antiferromagnets -- see for example R. W. Sanders, R. M. Belanger, M. Motokawa, and V. Jaccarino *Phys. Rev. B* 23 1190 (1981).
6. A review of nonreciprocal surface waves is given by R. E. Camley, *Surface Science Reports*, 7 103 (1987).
7. R. E. Camley, *Surface Science Reports* 7 103 (1987).
8. For an introductory summary, see Waguik S. Ishak and Kok-Wai Chang, *Hewlett-Packard J.* 10 (February 1985).
9. J. E. Sethares, *J. Appl. Phys.* 53 2646 (1982).
10. D. M. Bolle and S. H. Talisa, *IEEE Trans. Micro. The. Teck.* MTT-29, 916 (1981).
11. S. P. Vernon, R. W. Sanders, and A. R. King, *Phys. Rev. B* 17 1460 (1978).
12. K. W. Blazey, H. Rohrer, and R. Webster, *Phys. Rev. B*, 4 2287 (1971).
13. R. C. Ohomann and M. Tinkham, *Phys. Rev.*, 123 438 (1961).
14. For a brief description of damping terms, see A. H. Morrish, The Physical Principles of Magnetism, (Krieger Publishing Co., Malaban, Florida, 1983) pgs. 549-551.
15. R. L. Stamps and R. E. Camley, *Phys. Rev. B*, 35 1919 (1987).
16. R. E. Camley and D. L. Mills, *Phys. Rev. B* 26, 1280 (1982) and C. Shu and A. Caillé, *Solid State Comm.* 42 233 (1982).
17. Reference 14, pg 617.

18. R. E. Camley, N.E. Glass, and A. A. Maradudin, J. Appl. Phys. 53, 3170(1982).
19. M. Weber and D. L. Mills, Phys. Rev. B, 27 2698(1983).
20. D. L. Mills, Phys. Rev. B., 15 3097(1977).
21. L. Remer, E. Mohler, W. Grill and B. Luthi, Phys. Rev. B, 30 3277(1984).
22. P. Halevi in reference 1 page 249.
23. C. E. Patton, private communication.
24. M. G. Cottam and A. A. Maradudin, in Surface Excitations ed. by V. M. Agranovich and R. Loudon (Elsevier Science Pub. pages 1-194) 1984.
25. D. L. Mills, Phys. Rev. B 12 4036 (1975).
26. A. A. Maradudin and D. L. Mills, Phys. Rev. B 11 1892 (1975).
27. A. A. Maradudin and W. Zierau, Phys. Rev. B 14 484 (1976).
28. A. A. Maradudin and R. F. Wallis, J. Raman Spectroscopy, 10 85 (1981).
29. G. I. Stegeman, J. J. Burke and D. G. Hall, Opt. Lett., 8 383(1983).
30. C. W. Hsue and T. Tamir, J. Opt. Soc. Am. A, 1 923(1984).
31. T. Tamir, A. A. Oliner, Proc. IEE, 110 310(1963).
32. S. Rice, Comm. Pure. Appl. Math, 14 351(1951).
33. E. Kruger and E. Kretshmann, Z. Physik, 237 1(1970).
34. J. M. Elson and R. H. Ritchie, Phys. Rev. B, 4 4129(1971).
35. B. Laks, D.L. Mills and A. A. Maradudin, Phys. Rev. B, 23 4965(1981).
36. F. Toigo, A. Marvin, V. Celli and N. R. Hill, Phys. REv. B, 15 5618(1977).
37. G. S. Agarwal, Phys. REv. B, 14 846(1976).
38. N. E. Glass and A. A. Maradudin, J. Appl. Phys., 54 796(1983).
39. D. L. Mills and E. Burstein, Rep. Prog. Phys. 37 817 (1974).
40. R. E. Camley, Phys. Rev. Lett., 45 283 (1980).

APPENDICES

A. Fresnel Relations.

The amplitudes of the driving fields are written in terms of the incident wave's amplitude, $H^{(i)}$. The relations are derived in the usual manner by imposing Maxwell's continuity conditions on the unperturbed tangential \vec{H}^0 fields and the tangential \vec{E}^0 fields. Using

$$k_o^2 = k_{xo}^2 + k_{yo}^2 = \omega_o^2 \quad (A.1)$$

for the magnitude of the free space wavevector, and

$$k^2 = k_{xo}^2 + k_y^2 = \epsilon_2 \omega_o^2 (\mu_1^2 - \mu_2^2) / \mu_1 \quad (A.2)$$

for the magnitude of the wavevector in the material, the amplitudes of the components of the transmitted wave are

$$H_x^0(k_x, y) = \frac{2\epsilon_2 k_o k_{yo} k_y}{k^2 k_{yo} + \epsilon_2 k_o^2 k_y} H^{(i)} e^{iyk_y} \quad (A.3)$$

for the x component, and

$$H_y^o(k_x, y) = \frac{-2\epsilon_2 k_o k_{yo} k_{xo}}{k^2 k_{yo} + \epsilon_2 k_o^2 k_y} H^{(i)} e^{iyk_y} \quad (\text{A. 4})$$

for the y component.

Finally, the ratio of the transmittance to the reflectance is found to be

$$\frac{T}{R} = \left[\frac{2\epsilon_2 k k_o k_{yo}}{k^2 k_{yo} - \epsilon_2 k_o^2 k_y} \right]^2 \quad (\text{A.5})$$

B. Green's Functions

When the source points are in the material ($y' > 0$), the Green's functions satisfy inhomogeneous equations for $y > 0$ (in the material) given by

$$\begin{bmatrix} D^2 + \omega_o^2 \epsilon_2 \mu_1 & i(\omega_o^2 \epsilon_2 \mu_2 - k_x D) \\ -i(\omega_o^2 \epsilon_2 \mu_2 + k_x D) & -(k_x^2 - \omega_o^2 \epsilon_2 \mu_1) \end{bmatrix} \vec{g}(k_x; y, y') = -\frac{4\pi\epsilon_2}{c\epsilon_1} \vec{I} \delta(y - y') \quad (\text{B.1})$$

and homogeneous equations for $y < 0$ (outside the material) given by

$$\begin{bmatrix} D^2 + \omega_o^2 & -ik_x D \\ -ik_x D & \omega_o^2 - k_x^2 \end{bmatrix} \vec{g}(k_x; y, y') = 0 \quad (\text{B.2})$$

These equations can be uncoupled and solved for the homogeneous and particular solutions g_{ij} , just as in the case where $y' < 0$. Also, one can derive boundary conditions, in the manner described in chapter 5, that are identical to those of equations (5.34) and (5.43).

With the definition

$$C = -\gamma(\mu_1 \alpha \text{sgn}(y - y') + \mu_2 k_x) \quad (\text{B.3})$$

this prescription results in the following Green's functions for $y > 0$ and $y' > 0$:

$$g_{\alpha\alpha}^+ = \frac{2\pi A}{\alpha \omega_o^2 \epsilon_1 \mu_1} \left\{ \left[\frac{A-C}{A-B} \right] e^{-\alpha(y+y')} - e^{-\alpha|y-y'|} \right\} \quad (\text{B.4})$$

$$g_{xy}^+ = -\frac{2\pi A}{\alpha\omega_0^2\epsilon_1\mu_1} \left\{ \left[\frac{A-C}{A-B} \right] e^{-\alpha(y+y')} e^{-\alpha|y-y'|} \right\} \left\{ \omega_0^2\epsilon_2\mu_2 - \alpha k_x \operatorname{sgn}(y-y') \right\} \quad (\text{B.5})$$

$$g_{yy}^+ = \frac{2\pi i (\omega_0^2\epsilon_2\mu_2 - \alpha k_x \operatorname{sgn}(y-y'))}{\alpha\omega_0^2\epsilon_1\mu_1 A} \left\{ \omega_0^2\epsilon_2\mu_2 \left[e^{-\alpha|y-y'|} \left(\frac{A-C}{A-B} \right) e^{-\alpha(y+y')} \right] \right. \\ \left. - \alpha k_x \left[\operatorname{sgn}(y-y') e^{-\alpha|y-y'|} - \operatorname{sgn}(y+y') \left(\frac{A-C}{A-B} \right) e^{-\alpha(y+y')} \right] \right\} + \frac{4\pi\epsilon_2}{\epsilon_1 A} \delta(y-y') \quad (\text{B.6})$$

$$g_{yx}^+ = \frac{2\pi i}{\alpha\omega_0^2\epsilon_1\mu_1} \left\{ \omega_0^2\epsilon_2\mu_2 \left[\left(\frac{A-C}{A-B} \right) e^{-\alpha(y+y')} e^{-\alpha|y-y'|} \right] \right. \\ \left. + \alpha k_x \left[\operatorname{sgn}(y-y') e^{-\alpha|y-y'|} - \operatorname{sgn}(y+y') \left(\frac{A-C}{A-B} \right) e^{-\alpha(y+y')} \right] \right\} \quad (\text{B.7})$$

For $y < 0$ and $y' > 0$,

$$g_{xx}^- = \frac{4\pi AB}{\alpha\omega_0^2\epsilon_1\mu_1} \left(\frac{1}{A-B} \right) e^{\gamma y} e^{-\alpha y'} \quad (\text{B.8})$$

$$g_{xy}^- = \frac{-4\pi i B}{\alpha\omega_0^2\epsilon_1\mu_1} \left(\frac{1}{A-B} \right) (\omega_0^2\epsilon_2\mu_2 + \alpha k_x) e^{\gamma y} e^{-\alpha y'} \quad (\text{B.9})$$

$$g_{yy}^- = \frac{-4\pi B k_x}{\gamma \alpha \omega_o^2 \epsilon_1 \mu_1} \left(\frac{1}{A-B} \right) (\omega_o^2 \epsilon_2 \mu_2 + \alpha k_x) e^{\gamma y} e^{-\alpha y}$$

(B.10)

$$g_{yx} = \frac{-4\pi i A B k_x}{\alpha \gamma \omega_o^2 \epsilon_1 \mu_1} \left(\frac{1}{A-B} \right) e^{\gamma y} e^{-\alpha y}$$

(B.11)

Finally, as discussed in section 5.1, note that these forms are appropriate only for the case α and γ are real. When α or γ is imaginary, the appropriate forms are obtained by letting α go to $-i\alpha$ or γ go to $-i\gamma$ in the above expressions. Likewise, when damping is present and α and γ are complex, the correct transformation is α to $-i\alpha^*$ and γ to $-i\gamma^*$.

C. Landau-Lifshitz damping terms.

The form of the Landau damping terms which appear in the equations of motion can be obtained from the free energy of the antiferromagnet. Using the definitions of the effective fields acting on each sublattice (equations 2.2 and 2.3), the free energy is written as the sum of the free energies of each sublattice:

$$F = -\vec{M}_a \cdot \vec{H}_a - \vec{M}_b \cdot \vec{H}_b \quad (C.1)$$

In the calculation of the susceptibilities only terms to first order in the transverse magnetizations M_x and M_y were kept. The same approximation of small precession angles in the free energies must be made to second order in M_x and M_y for consistency. A Taylor expansion of M_z about $M_x=0$ and $M_y=0$ yields the relation

$$M_z^a = M - \frac{1}{2M} \left[(M_x^a)^2 + (M_y^a)^2 \right] \quad (C.2)$$

for the z component of the A sublattice magnetization and

$$M_z^b = -M + \frac{1}{2M} \left[(M_x^b)^2 + (M_y^b)^2 \right] \quad (C.3)$$

for the z component of the B sublattice magnetization. M is the saturation magnetization.

These expressions are substituted into the free energies and only terms to second order in the transverse magnetizations are kept. In the small precession angle approximation, there are no forces in the z direction. Since the gradient of the free energy with respect to \vec{M}

gives the effective forces acting on the magnetizations, the gradient in the M_z direction vanishes. The transverse gradient for the A sublattice is defined as

$$\nabla_a = \hat{x} \frac{\partial}{\partial M_x^a} + \hat{y} \frac{\partial}{\partial M_y^a} \quad (\text{C.4})$$

The transverse gradient for the B sublattice, ∇_b , is defined similarly. Taking $\nabla_a F$, one finds:

$$\nabla_a F = \hat{x} \left[\lambda(M_x^a + M_x^b) + (H_0 + h_a) \frac{M_x^a}{M} - h_x \right] + \hat{y} \left[\lambda(M_y^a + M_y^b) + (H_0 + h_a) \frac{M_y^a}{M} - h_y \right] \quad (\text{C.5})$$

h_a and H_0 are the anisotropy and applied fields, respectively. With C.5, the following equality can be verified:

$$\nabla_a F = \frac{1}{M^2} \vec{M}_a \times (\vec{M}_a \times \vec{H}_a) \quad (\text{C.6})$$

A similar expression can be obtained for the B sublattice magnetizations. The Bloch equations of motion with Landau damping can thus be written

$$\dot{\vec{M}}_a = \gamma \vec{M}_a \times \vec{H}_a - \Lambda \nabla_a F \quad (\text{C.7})$$

and

$$\dot{\vec{M}}_b = \gamma \vec{M}_b \times \vec{H}_b - \Lambda \nabla_b F \quad (\text{C.8})$$

**SODIUM LIDAR STUDIES OF THE HORIZONTAL VARIABILITY OF
GRAVITY WAVES IN THE MESOSPHERE**

BY

KANG HYON KWON

**B.S., University of Illinois, 1984
M.S., University of Illinois, 1986**

THESIS

**Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1989**

Urbana, Illinois

ABSTRACT

A data analysis technique for determining gravity wave intrinsic parameters is presented. The intrinsic parameters include the horizontal and vertical wavelengths, period, and wave propagation direction. The technique involves measuring the altitude variations of the wave induced density perturbations in the mesospheric Na layer. The intrinsic horizontal wavelength of a wave observed with an airborne lidar in November 1986 was estimated to be about 85 km, and the vertical wavelength was 4.1 km. The intrinsic period was about 1.7 hours, and the propagation direction was almost due south.

Kinetic energy horizontal and vertical wavenumber spectra of horizontal winds are inferred from Na density profiles collected with the airborne lidar during two flights in November 1986. The two flights include one roundtrip from Denver, Colorado to Springfield, Illinois and another roundtrip from Denver to the Pacific Coast. The horizontal wavenumber spectra exhibited an approximately $k_x^{-1.2}$ dependence at horizontal scales from 70 to 700 km, while the vertical wavenumber spectra exhibited an approximately $k_z^{-2.7}$ dependence at vertical scales from 2 to 10 km.

In November 1986, joint lidar/radar observations were conducted. The lidar observations include the airborne observations and ground-based observations at Broomfield and Denver, Colorado. The radar observations were obtained at Platteville, Colorado with an ST radar. These joint observations revealed that waves with periods of approximately 6 hours and 2 hours were dominant at the altitudes that correspond to the bottomside of the Na layer.

The characteristics of sporadic Na layers observed at Mauna Kea Observatory, Hawaii (20°N, 155°W) and at Nordlysstasjonen, Svalbard, Norway (78°N, 16°E) are also described. The layers observed in Hawaii formed either in the late evening or in the early morning. The mean time of the maximum peak density of the early morning layers occurred about 6 hours after that of the late evening layers. The mechanisms responsible for creating these layers appear to be related to diurnal tides and sporadic E layers.

DEDICATION

To my father and mother

ACKNOWLEDGMENT

I wish to express my sincere appreciation to Professor Chester S. Gardner under whose supervision this work was performed. His guidance, friendship, and inspiration have been invaluable to me. I would also like to thank Professor C. H. Liu for many helpful suggestions during the course of my academic progress. I thank Professor Steve Franke for his comments and for serving on my doctoral committee, and I thank my fellow graduate students at the Electro-optics Systems Laboratory for their encouragement and assistance.

I would like to specially thank my wife Casey for her unending support and encouragement.

TABLE OF CONTENTS

	Page
1. INTRODUCTION.....	1
2. AIRBORNE SODIUM LIDAR MEASUREMENTS OF GRAVITY WAVE INTRINSIC PARAMETERS.....	5
2.1 Introduction.....	5
2.2 Estimation of Gravity Wave Intrinsic Parameters.....	6
2.3 Experimental Data.....	18
2.4 Summary.....	32
3. AIRBORNE SODIUM LIDAR OBSERVATIONS OF HORIZONTAL AND VERITCAL WAVENUMBER SPECTRA OF MESOPAUSE DENSITY AND WIND PERTURBATIONS.....	33
3.1 Introduction.....	33
3.2 Layer Density Response.....	34
3.3 Description of the Experiment.....	41
3.4 Results.....	44
3.5 Discussion.....	79
3.6 Summary.....	82
4. CORRELATIVE RADAR AND AIRBORNE SODIUM LIDAR OBSERVATIONS OF THE VERTICAL AND HORIZONTAL STRUCTURE OF GRAVITY WAVES AND TIDES NEAR THE MESOPAUSE.....	84
4.1 Introduction.....	84
4.2 Description of the Experiment.....	85
4.2.1 Radar.....	85
4.2.2 Lidar.....	86
4.3 Radar Observations.....	86

	Page
4.4 Ground-based Lidar Observations.....	91
4.5 Airborne Lidar Observations.....	109
4.6 Summary.....	112
5. LIDAR OBSERVATIONS OF SPORADIC SODIUM LAYERS AT MAUNA KEA OBSERVATORY, HAWAII.....	115
5.1 Introduction.....	115
5.2 Observations.....	116
5.3 Discussion.....	134
5.4 Summary.....	139
6. LIDAR OBSERVATIONS OF THE SODIUM LAYER AT SVALBARD, NORWAY.....	143
6.1 Overviews.....	143
6.2 Lidar Observations at Svalbard in July and September, 1986.....	147
6.3 Lidar Observations at Svalbard in November, 1987 and January, 1988.....	155
7. CONCLUSIONS AND RECOMMENDATIONS.....	167
7.1 Conclusions.....	167
7.2 Recommendations for Future Work.....	170
APPENDIX I. RMS ERRORS IN GRAVITY WAVE INTRINSIC PARAMETERS FOR AIRBORNE LIDAR OBSERVATIONS OVER A CIRCULAR FLIGHT PATH.....	172
APPENDIX II. RMS ERRORS IN GRAVITY WAVE INTRINSIC PARAMETERS FOR MULTIPLE GROUND-BASED LIDAR OBSERVATIONS.....	175
APPENDIX III. CHARACTERISTICS OF SPORADIC SODIUM LAYERS OBSERVED AT MAUNA KEA OBSERVATORY, HAWAII.....	178

	Page
REFERENCES.....	182
VITA.....	188

LIST OF TABLES

Tables	Page
2.1 The Parameters of the UTUC Na Lidar System and NCAR Electra Aircraft...	19
2.2 Summary of the Flights and System Performance.....	21
2.3 Intrinsic Parameters of the Dominant Wave Observed During the Westward Flight on November 17-18, 1986.....	27
3.1 Horizontal Velocities of the Maxima and Minima of the Centroid Height of the Na Layer Observed During the Eastward Flight on November 15-16, 1986.....	55
3.2 Monochromatic Wave Parameters Estimated from the Horizontal Wavenumber Spectra for the Eastward Flight on November 15-16, 1986.....	61
3.3 Summary of the Horizontal and Vertical Wavenumber Spectra Parameters Measured on the Eastward and Westward Flights.....	74
4.1 Summary of the Airborne and Ground-based Na Lidar Observations.....	87
5.1 Statistics of Na Sporadic Layers Observed at Mauna Kea, Hawaii.....	117
5.2 Comparison of Sporadic Na Layers Observed in the Late Evening and Early Morning at Mauna Kea, Hawaii.....	132
5.3 Characteristics of the Most Prominent Sporadic Na Layers Reported in Literature.....	133
6.1 Sodium Lidar Observation Times at Svalbard (1987-1988).....	144
6.2 Sodium Layer Parameters Measured at Nordlysstasjonen (78°12'N, 15°15'E), Svalbard During the July and September 1987 Campaigns.....	153

LIST OF FIGURES

Figures	Page
2.1 The relationship between the horizontal components of the wave propagation direction and background atmospheric wind.....	9
2.2 Ground track of the circular flight path with radius R.....	13
2.3 Configuration for three ground-based lidars located at the corners of an equilateral triangle with sides of length R.....	16
2.4 Ground tracks of the three flights of the airborne Na lidar experiment in November, 1986. The flights were conducted out of Denver, Colorado.....	20
2.5 Sodium density profiles collected during the westbound and eastbound legs of the westward flight on November 17-18, 1986. The profiles have been filtered vertically with a cutoff of 3 km and horizontally with a cutoff of 50 km. The profiles have been also normalized so that each has the same column abundance, and are plotted on a linear scale.....	22
2.6 The altitude variations of a local Na density maximum and minimum measured during the westward flight on November 17-18, 1986. Circles represent the altitudes of the density maximum, and crosses represent the altitudes of the density minimum. The solid lines represent the least-squares fitted altitudes of the density maximum and minimum.....	24
2.7 Average vertical wavenumber power spectrum of the Na density profiles collected from 0131 to 0138 MST ¹ during the westward flight on November 17-18, 1986.....	25
2.8 a) Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound leg of the westward flight on November 17-18, 1986. The data were filtered vertically with a cutoff of 1 km before the spectra were computed.....	29

Figures	Page
2.8 b) Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the eastbound leg of the westward flight on November 17-18, 1986. The data were filtered vertically with a cutoff of 1 km before the spectra were computed.....	30
3.1 A photograph of the NCAR Electra aircraft.....	42
3.2 A photograph of the interior of the NCAR Electra showing the lidar installation.....	43
3.3 Sodium density profiles collected during the eastbound leg of the eastward flight on November 15, 1986. The profiles were filtered vertically with a cutoff of 4 km and horizontally with a cutoff of 70 km. The density profiles were also normalized so that each had the same column abundance and were plotted on a linear scale.....	45
3.4 Vertical and longitudinal variations of the relative density perturbations computed from the Na lidar data collected during the eastbound leg of the eastward flight on November 15, 1986. Before the data were computed, the density profiles were filtered vertically with a cutoff of 1 km and horizontally with a cutoff of 70 km. Then the relative density perturbations were computed and filtered horizontally with a cutoff of 180 km.....	46
3.5 Horizontal wavenumber spectrum for the eastbound leg of the eastward flight on November 15, 1986. The data were filtered vertically with a cutoff of 1 km before the spectrum was computed. The straight line is a linear regression fit which was used to estimate the spectral slope over horizontal scales from 70 to 700 km.....	47
3.6 Vertical wavenumber spectrum for the eastbound leg of the eastward flight on November 15, 1986. The data were filtered horizontally with a	

Figures	Page
cutoff of 70 km before the spectrum was computed. Then the shot noise level was estimated and subtracted from the spectrum before the spectral slope was estimated. The straight line is a linear regression fit which was used to estimate the spectral slope over vertical scales from 2 to 10 km.....	49
3.7 Sodium density profiles collected during the westbound leg of the eastward flight on November 15-16, 1986. The profiles were processed in the same manner as those of the eastbound leg of the eastward flight plotted in Figure 3.3.....	50
3.8 Vertical and longitudinal variations of the relative density perturbations computed from the Na lidar data collected during the westbound leg of the eastward flight on November 15-16, 1986. To compute these relative density perturbations, the Na profiles were first filtered vertically with a cutoff of 1 km and horizontally with a cutoff of 70 km. Then the relative density perturbations were computed and filtered horizontally with a cutoff of 233 km.....	51
3.9 Horizontal wavenumber spectrum for the westbound leg of the eastward flight on November 15-16, 1986. The data were processed in the same manner as the data of the eastbound leg of the eastward flight.....	53
3.10 The longitudinal variations of the centroid height of the Na layer observed during the eastward flight on November 15-16, 1986.....	54
3.11 The horizontal wavelengths of the waves observed with the ground-based (after <i>Gardner and Voelz</i> , 1987) and airborne Na lidars are plotted versus period. Crosses represent the ground-based observations obtained at Urbana, Illinois during the winter seasons from 1980 to 1986. Circles represent the intrinsic zonal wavelengths and intrinsic periods measured	

Figures	Page
<p>during the eastward flight on November 15-16, 1986. The straight line is a maximum likelihood linear regression fit which was used to estimate the slope for the ground-based observations.....</p>	62
<p>3.12 The kinetic energy of quasi-monochromatic gravity waves plotted versus temporal frequency. Crosses represent the ground-based Na lidar observations obtained at Urbana, Illinois during the winter seasons from 1980 to 1986 (after <i>Gardner and Voelz</i>, 1987). Circles represent the kinetic energy distribution for the quasi-monochromatic waves measured during the eastward flight on November 15-16, 1986. The shaded circle is the data measured during the westbound leg, and the open circles are the data measured during the eastbound leg. The straight line represents a maximum likelihood linear regression fit for the ground-based measurements.....</p>	63
<p>3.13 Vertical wavenumber spectrum for the westbound leg of the eastward flight on November 15-16, 1986.....</p>	65
<p>3.14 a) Horizontal wavenumber spectrum for the eastbound leg of the westward flight on November 17-18, 1986.....</p>	66
<p>3.14 b) Horizontal wavenumber spectrum for the westbound leg of the westward flight on November 17-18, 1986.....</p>	67
<p>3.15 a) Vertical wavenumber spectrum for the eastbound leg of the westward flight on November 17-18, 1986.....</p>	68
<p>3.15 b) Vertical wavenumber spectrum for the westbound leg of the westward flight on November 17-18, 1986.....</p>	69
<p>3.16 The rms horizontal wind velocities inferred from the airborne Na lidar data collected during the eastward (November 15-16, 1986) and westward</p>	

Figures	Page
(November 17-18, 1986) flights.....	70
3.17 The rms horizontal wind velocities inferred from the airborne Na lidar data, and from the ground-based Na lidar data obtained at Mauna Kea Observatory (MKO), Hawaii (20°N,155°W), NCAR in Broomfield, Colorado (40°N, 105°W), UIUC, Illinois (40°N, 88°W), and Goddard Space Flight Center (GSFC), Maryland (39°N, 78°W). The dots and lines for the measurements at MKO, NCAR, and GSFC indicate the average values and the ranges of the measured values, respectively. The dot for the UIUC data represents a regression fitted value for the middle of November. The regression fit was performed over the data obtained from 1984 to 1986. The error bar for the UIUC data indicates the rms difference between the regression fit and the measured values.....	72
3.18 a) Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the eastbound leg of the eastward flight on November 15-16, 1986.....	75
3.18 b) Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound leg of the eastward flight on November 15-16, 1986.....	76
3.18 c) Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the eastbound leg of the westward flight on November 17-18, 1986.....	77
3.18 d) Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound leg of the westward flight on November 17-18, 1986.....	78
4.1 Vertical profiles of a) the meridional component and b) the zonal	

Figures	Page
component of average background wind velocity measured with Platteville ST radar during the period from November 4 to 20, 1986. The data were provided by the University of Colorado.....	88
4.2 Diurnal variation of the horizontal wind perturbations measured with the Platteville ST radar during the period from November 4 to 20, 1986. The boxes represent the hourly velocity measurements, and the solid curve represents a sinusoidal least-square fit. The data were provided by the University of Colorado.....	89
4.3 Vertical profiles of a) the meridional component and b) the zonal component of the 6-hour period horizontal winds. The data were provide by the University of Colorado.....	90
4.4 Sodium density profiles measured on the night of November 19-20, 1986. The profiles have been filtered vertically with a cutoff of 3 km and temporally with a cutoff of 50 min. The profiles have been normalized so that each has the same column abundance, and are plotted on a linear scale at 10 min intervals.....	92
4.5 Temporal variations of the layer centroid height, rms width, and column abundance measured on November 19-20, 1986.....	93
4.6 a) Temporal frequency spectrum computed for the whole layer from data collected on November 19-20, 1986. The Na density profiles were vertically filtered with a cutoff of 1 km before the spectra were computed. The straight line is a linear regression fit which was used to estimate the spectral slopes over temporal scales from 30 to 360 min.....	95
4.6 b) Temporal frequency spectra computed for the layer bottomside and topside from data collected on November 19-20, 1986. The Na density profiles	

Figures	Page
<p>were vertically filtered with a cutoff of 1 km before the spectra were computed. The straight lines are linear regression fits which were used to estimate the spectral slopes over temporal scales from 30 to 360 min.....</p>	96
4.7 Relative temporal variations of the Na density from 81 to 92 km measured on November 19-20, 1986. The Na density profiles were vertically filtered with a cutoff of 5 km and temporally with a cutoff of 90 min. The diagonal lines indicate apparent vertical phase progressions. The estimated phase velocity is 1.8 m s^{-1}	97
4.8 The vertical wind velocity estimated at the altitude of 82 km. The Na density profiles were initially filtered vertically with a cutoff of 5 km and temporally with a cutoff of 60 min.....	99
4.9 The vertical wind velocity estimated at the altitude of 101 km. The Na density profiles were initially filtered vertically with a cutoff of 5 km and temporally with a cutoff of 60 min.....	100
4.10 The average vertical wavenumber power spectrum computed for the Na density profiles collected from 2300 to 0100 MST on November 19-20, 1986.....	101
4.11 The rms horizontal wind velocity inferred from the Na density profiles collected on November 19-20, 1986. The profiles were initially filtered vertically with a cutoff of 1 km and temporally with a cutoff of 20 min.....	102
4.12 Temporal variations of the altitudes of local Na density maxima measured on November 19-20, 1986. The Na density profiles were initially filtered vertically with a cutoff of 3 km and temporally with a cutoff of 50 min.....	105
4.13 The average vertical wavenumber power spectrum computed for the Na density profiles collected from 0300 to 0625 MST on November 20, 1986..	107

Figures	Page
4.14 Vertical wavenumber spectrum measured on November 19-20, 1986. The Na density profiles were initially filtered temporally with a cutoff of 20 min. The straight line is a linear regression fit which was used to estimate the spectral slope over vertical scales ranging from 2 to 10 km.....	108
4.15 a) Longitudinal variations of the rms width of the Na layer observed during the eastward flight on November 15-16, 1986.....	110
4.15 b) Longitudinal variations of the column abundance of the Na layer observed during the eastward flight on November 15-16, 1986.....	111
4.16 The average vertical wavenumber power spectrum computed for the Na density profiles collected during the westbound leg of the eastward flight in the longitude range from 96°W to 102°W.....	113
5.1 Sodium density profiles measured at a) 2133 LST, b) 2218 LST, and c) 2224 LST at Mauna Kea Observatory on January 21, 1987.....	118
5.2 Sodium density profile measured at 2306 LST on January 21, 1987 at Mauna Kea Observatory.....	120
5.3 Temporal variation of the altitude of the dominant sporadic Na layer observed on January 21-22, 1987 at Mauna Kea Observatory. The straight line is a linear regression fit which was used to estimate the average vertical velocity.....	121
5.4 Sodium density profile measured at 0407 LST on January, 22, 1987 at Mauna Kea Observatory.....	122
5.5 Temporal variation of the density at the peak of the dominant sporadic Na layer observed on January 21-22, 1987 at Mauna Kea Observatory.....	123
5.6 Temporal variations of the widths of the dominant sporadic Na layer observed on January 21-22, 1987 at Mauna Kea Observatory.....	125

Figures	Page
5.7 Temporal variation of the column abundance measured on January 21-22, 1987 at Mauna Kea Observatory.....	126
5.8 Temporal variations of the peak density and altitude of the high altitude sporadic Na layer observed on January 21, 1987 at Mauna Kea Observatory.....	127
5.9 Temporal variation of the altitude of the short-lived sporadic Na layer observed on January 21, 1987 at Mauna Kea Observatory. Crosses represent the four measurement points.....	129
5.10 Temporal variations of the altitudes of the sporadic Na layers measured from January 18 to 22, 1987 at Mauna Kea Observatory. Dots represent the times and altitudes of the maximum peak densities. Squares, triangles, circles, and diamonds represent starting times and altitudes. The sporadic layers are numbered in accordance with Appendix 3.....	130
5.11 Sodium density profiles measured at 0256 LST on January 18, 0301 LST on January 20, and 0255 LST on January 22, 1987.....	136
5.12 Temporal variations of the altitudes of local Na density maxima on January 21-22, 1987 at Mauna Kea Observatory. The measurement times of the density profiles of Figures 5.1b), 5.1c), 5.2, and 5.4 are marked on the bottom of this figure.....	138
5.13 Temporal variations of the altitudes of local Na density maxima and minima from 2100 to 2330 LST on January 21, 1987 at Mauna Kea Observatory. The solid curves represent the altitudes of the sporadic layers, the dashed curves the local density maxima, and the dotted curves the local density minima. The sporadic layers are numbered in accordance with Table 5.2. The measurement times of the density profiles of Figures 5.1a), 5.1b),	

Figures	Page
5.1c), and 5.2 are marked on the bottom of this figure.....	140
6.1 Seasonal variations of Na column abundance measured a) at Svalbard, Norway (78°N) and b) at Urbana, Illinois (40°N). Symbols denote average values for the observation period, while lines denote range of values.....	145
6.2 Seasonal variations of Na layer centroid height measured a) at Svalbard, Norway (78°N) and b) at Urbana, Illinois (40°N). Symbols denote average values for the observation period, while lines denote range of values.....	146
6.3 Seasonal variations of Na layer rms width measured a) at Svalbard, Norway (78°N) and b) at Urbana, Illinois (40°N). Symbols denote average values for the observation period, while lines denote range of values.....	148
6.4 Average Na density profile above Nordlysstasjonen, Svalbard during the period July 10-15, 1987. The data were smoothed using modified Hamming window with FWHM = 6 km.....	150
6.5 Sodium density profile above Nordlysstasjonen, Svalbard at 2152 UT September 7, 1987. The integration period was 14 min, and the data were spatially smoothed using a low-pass filter with a cutoff frequency of 0.6 km ⁻¹	151
6.6 Sodium density profile above Nordlysstasjonen, Svalbard at midnight September 9, 1987. The integration period was 40 min, and the data were spatially smoothed using a low-pass filter with a cutoff frequency of 0.6 km ⁻¹	152
6.7 a) Temporal variations and b) temporal power spectrum of the Na column	

Figures	Page
abundance measured during the 57-hour period starting at 1500 LST on November 7, 1987.....	156
6.8 a) Temporal variations and b) temporal power spectrum of the Na layer centroid height measured during the 57-hour period starting at 1500 LST on November 7, 1987.....	157
6.9 a) Temporal variations and b) temporal power spectrum of the Na rms width measured during the 57-hour period starting at 1500 LST on November 7, 1987.....	158
6.10 a) Temporal variations and b) temporal power spectrum of the vertical winds at the altitude of 98 km. The vertical winds were inferred from the temporal variations of the Na density gradients on the layer topside measured during the 57-hour period starting at 1500 LST on November 7, 1987.....	159
6.11 a) Temporal variations and b) temporal power spectrum of the Na column abundance measured during the 72-hour period starting at local midnight on January 10, 1988.....	160
6.12 a) Temporal variations and b) temporal power spectrum of the Na centroid height measured during the 72-hour period starting at local midnight on January 10, 1988.....	161
6.13 a) Temporal variations and b) temporal power spectrum of the Na rms width measured during the 72-hour period starting at local midnight on January 10, 1988.....	162
6.14 a) Temporal variations and b) temporal power spectrum of the vertical winds at the altitude of 98 km. The vertical winds were inferred from the temporal variations of the Na density gradients on the layer topside	

Figures	Page
measured during the 72-hour period starting at local midnight on January 10, 1988.....	163
6.15 Examples of sporadic Na layers observed at Svalbard a) on November 9, 1987 and b) on January 10, 1988.....	165
6.16 a) A bifurcated sporadic Na layer observed at Svalbard on November 9, 1988, and b) the temporal evolution of the sporadic Na layer.....	166

1. INTRODUCTION

The atmospheric Na layer is generally confined to an altitude region between 80 and 110 km, which is the transition region from the mesosphere and thermosphere. The mesopause characterizes the boundary between the mesosphere and thermosphere, and is typically the coldest region of the atmosphere. The altitude of the mesopause is nearly 90-95 km, and the maximum Na density occurs near the same altitude. Because this region is relatively inaccessible to in-situ measurement techniques, it has remained as the least understood region of the earth's atmosphere. For the past two decades, Na lidar techniques have been developed to explore this region. In recent years, these Na lidar techniques have matured to the extent that they are now providing very important information on the dynamics and temperature structure of this region.

Studies of the Na layer began when *Slipher* [1929] discovered nighttime spectral emission at the resonance wavelength of the Na D₂ line of 589 nm, which is the result of the relaxation of the excited atmospheric Na atoms to neutral atoms. In the 1950s and 1960s, observations of resonantly scattered sunlight at the Na wavelength defined the diurnal and seasonal variations in column abundance [*Blamont and Donahue*, 1961; *Gadsen and Purdy*, 1970; *Burnett et al.*, 1975], and rocketborne dayglow measurements discovered the sharp upper and lower boundaries of the layer [*Hunten and Wallace*, 1967; *Donahue and Meier*, 1967].

Lidar observations of the Na layer were first made in England following the invention of the tunable dye laser in the late 1960s [*Bowman et al.*, 1969]. Because of the excellent temporal and vertical resolution, lidar observations have revealed detailed information about the layer. The vertical density profile of the layer resembles approximately the Gaussian distribution shape with a peak density of 10^3 - 10^4 cm⁻³ near 90-95 km, and the width is about 10 km Full Width at Half Maximum (FWHM). The major source of Na is now believed to be meteoric ablation, and the major sink mechanism is chemical reaction on the bottomside including the neutral reaction $\text{Na} + \text{O}_2 + \text{M} \rightarrow \text{NaO}_2 + \text{M}$ [*Swider*, 1985].

The first lidar study of the seasonal variations of the Na layer was made by *Gibson and Sandford* [1971] at Winkfield, England (51°N, 1°W). Similar studies have also been conducted by *Megie and Blamont* [1977] at Haute Provence, France (44°N, 6°E), by *Simonich et al.* [1979] at Sao Paulo, Brazil (23°S, 46°W), by *Gardner et al.* [1986] at Urbana, IL (40°N, 88°W), and by *Tilgner and von Zahn* [1988] at Andoya, Norway (69°N, 16°E). The Na column abundance at mid-latitudes in the Northern Hemisphere varies from a summer minimum of about 3×10^9 cm⁻² to a winter maximum of about 10^{10} cm⁻² in December and January. The seasonal and geographical variations in Na abundance at mid- and low- latitudes are now believed to be related to changes in the mesopause temperature that affect the reaction rates of the main chemical loss processes for Na [*Swider*, 1985; *Jegou et al.*, 1985b].

High resolution temperature profiles have been obtained also in recent years by lidar probing of the hyperfine structure of the Na D₂ line [*Fricke and von Zahn*, 1985; *von Zahn and Neuber*, 1987]. The Na resonance line is actually a Doppler broadened doublet whose shape is a strong function of temperature. Temperature can be inferred by using a narrowband frequency scanning lidar to measure the Na line shape.

Gravity waves and tides are now widely recognized to play a major role in determining the large-scale circulation and structure of the middle atmosphere. Gravity waves are believed to be the major source of energy transport between the lower and upper atmosphere. Turbulence caused by breaking waves has a significant influence on the transport of minor species and on the thermal and density structure of the mesosphere and the thermosphere [*Lindzen*, 1981; *Holton*, 1982 and 1983; *Fritts et al.*, 1984; *Fritts*, 1984].

The Na layer has proven to be an excellent tracer of wave motions. Early lidar measurements of the Na profile did reveal wavelike perturbations that were attributed to acoustic gravity waves [*Rowlett et al.*, 1978; *Juramy et al.*, 1981]. But it was not until the work of *Chiu and Ching* [1978] and *Gardner and Shelton* [1985] that the interaction of gravity waves with the layer was sufficiently well understood that the wave parameters could be inferred from lidar

measurements of the Na profile. The most extensive lidar study of gravity wave events was recently published by *Gardner and Voelz* [1987]. During 34 nights of measurements at Urbana, IL between December 1980 and May 1986, a total of 171 monochromatic gravity waves were observed in the Na layer. From these measurements, *Gardner and Voelz* [1987] found surprisingly systematic relationships between horizontal and vertical wavelengths and the observed periods of the waves. Studies of tides using Na lidars have been also reported by *Batista et al.* [1985] and *Kwon et al.* [1987].

Because of the influence of background atmospheric wind field, wave parameters which are estimated from most ground-based lidar observations are not the intrinsic parameters. In addition, because most lidar observations have been made at fixed elevation angles, the studies of gravity waves and tides are usually limited to exploring only the vertical and temporal structures of the waves. The horizontal structure of the waves is then inferred from the vertical and temporal observations. Also very little is known about the geographical distribution of wave activity.

In order to study the geographical distribution and horizontal structure of the waves, the University of Illinois at Urbana-Champaign (UIUC) group has conducted several Na lidar campaigns from 1986 to 1988. The campaigns include an airborne campaign which was conducted in November, 1986 with the support of the National Center for Atmospheric Research - Research Aviation Facility (NCAR-RAF) in Broomfield, Colorado. A total of three flights were made over the Great Plains, Rocky Mountains, and Pacific Coast. Additional campaigns were conducted at Broomfield, Colorado in November, 1986, at Mauna Kea Observatory, Hawaii in January, 1987, and at Nordlysstasjonen, Svalbard, Norway from June, 1987 to April, 1988. In order to process the airborne data, new data analysis techniques have been also developed. In this thesis, the new data analysis techniques and the results of the campaigns will be presented.

In Chapter 2, a data analysis technique for determining gravity wave intrinsic parameters will be presented. The intrinsic parameters include the horizontal and vertical wavelengths, period, and wave propagation direction. The technique is successfully applied to the airborne data collected during a roundtrip flight from Denver, Colorado to the Pacific Coast in November, 1986.

In Chapter 3, a method is presented for estimating kinetic energy horizontal and vertical wavenumber spectra of horizontal winds from the airborne data. These spectra are compared with spectra obtained from ground-based lidar, radar, shuttle re-entry, and Global Atmospheric Sampling Program observations. The rms horizontal wind velocities inferred from the airborne data are also compared with those inferred from ground-based lidar observations conducted in Hawaii, Colorado, Illinois, and Maryland.

In November, 1986, the University of Colorado group operated an ST radar in Platteville, Colorado in conjunction with the airborne lidar observations and ground-based lidar observations at Broomfield and Denver, Colorado. The results of the joint radar/lidar observations are presented in Chapter 4.

In January, 1987, the UIUC group installed and operated the Na lidar at the low-latitude site of Mauna Kea Observatory (20°N, 155°W), Hawaii. The Na layer occasionally exhibited sporadic developments of very dense narrow Na layers. The characteristics of these sporadic Na layers will be discussed in Chapter 5. The characteristics of these layers are also compared with the characteristics of similar layers observed at Sao Paulo, Brazil (23°S, 46°W) by *Clemesha et al.* [1978] and at Andoya, Norway (69°N, 16°E) by *von Zahn and Hansen* [1988].

From July 1987 to April 1988, a total of five Na lidar campaigns were conducted at Nordlysstasjonen, Svalbard, Norway (78°N, 15°E). The characteristics of the Na layer observed at this high latitude site of Nordlysstasjonen are discussed in Chapter 6.

2. AIRBORNE SODIUM LIDAR MEASUREMENTS OF GRAVITY WAVE INTRINSIC PARAMETERS

2.1 Introduction

For almost 20 years, resonance fluorescence lidar systems have been used to study the mesospheric metals. Because of its relatively high density and large resonant backscattering cross section, Na is the easiest metallic species to measure with lidar techniques and has been explored extensively since the late 1960s. The Na layer is generally confined to the region between 80 and 110 km with a peak near the mesopause at 90-95 km, where the density ranges from about 10^3 - 10^4 cm⁻³. The layer is an excellent tracer of wave motions, and Na lidar observations are now making important contributions to the understanding of gravity wave and tidal dynamics in the mesopause region [Batista *et al.*, 1985; Gardner and Voelz, 1987; Kwon *et al.*, 1987]. Recent lidar-based studies of gravity waves have focused on characterizing quasi-monochromatic waves. For example, by analyzing a total of 171 quasi-monochromatic waves observed at Urbana, Illinois, Gardner and Voelz [1987] found surprisingly systematic relationships between horizontal and vertical wavelengths and the observed periods of the waves. Several radar-based studies of quasi-monochromatic waves have been also reported by Meek *et al.* [1985], Reid and Vincent [1987], and Manson and Meek [1988].

Because of the influence of the background atmospheric wind field, wave parameters which are estimated from most ground-based lidar observations are not the intrinsic parameters. In order to measure the intrinsic wave parameters and to explore the horizontal structure of gravity waves, the UIUC group has conducted several airborne Na lidar experiments. The first airborne observations were conducted in March 1983 on a single roundtrip flight from NASA Wallops Flight Facility to Albany, New York [Segal *et al.*, 1984]. This experiment demonstrated the feasibility of airborne measurements and revealed evidence of wave induced horizontal structure over the 650 km flight path. In November of 1986, the UIUC group conducted a second airborne Na lidar campaign with the support of the NCAR-RAF in Broomfield, Colorado. A total of three flights were made over the Great Plains, Rocky

Mountains, and Pacific Coast. The flights were conducted out of Stapleton International Airport in Denver, Colorado using the NCAR Electra aircraft.

In this chapter, a new data analysis technique for determining gravity wave intrinsic parameters including wave propagation direction will be presented. In Section 2.2, the theoretical basis for estimating the intrinsic wave parameters is discussed. Several examples for airborne lidar and multiple ground-based lidar experiments using this technique are also presented. In Section 2.3, the theoretical results are successfully applied to airborne Na lidar data obtained during a roundtrip flight from Denver to the Pacific Coast on November 17-18, 1986.

2.2 Estimation of Gravity Wave Intrinsic Parameters

Recently, *Gardner and Shelton* [1985] and *Gardner and Voelz* [1987] derived expressions for the density response of the atmospheric Na layer perturbed by a monochromatic gravity wave, and have used the expressions to determine the wave parameters. By retaining only the fundamental and first order perturbation terms, *Gardner and Voelz* [1987] have shown the density response of the Na layer perturbed by a wave to be

$$n_s(\underline{r},t) \approx n_o(\underline{r},t) - \left[n_o(\underline{r},t) + \gamma H \frac{\partial n_o(\underline{r},t)}{\partial z} \right] \frac{A e^{\beta z}}{\gamma - 1} \cos[\theta(\underline{r},t)] \quad (2.1)$$

where $n_s(\underline{r},t)$ = perturbed Na density,

$n_o(\underline{r},t)$ = Na density in the absence of the wave activity,

γ = ratio of specific heats (≈ 1.4),

H = atmospheric scale height (≈ 6 km),

$A e^{\beta z}$ = amplitude of the atmospheric density perturbations due to the gravity wave,

β = amplitude growth factor,

$\theta(\underline{r},t)$ = wave phase,

\underline{r} = $x\hat{x} + y\hat{y} + z\hat{z}$ = position vector,

- x = east-west component of the geocentric coordinates,
 y = north-south component of the geocentric coordinates, and
 z = vertical component of the geocentric coordinates.

The first term on the right-hand side of Equation (2.1) is the unperturbed layer profile, and the second term is the first-order perturbation term. When the wave phase, $\theta(r, t)$, is either 0 or π , the magnitude of the perturbation term reaches a maximum, and the perturbations result in a local maximum or minimum in the Na density profile. The wave phase can be written as

$$\theta(r, t) = \omega t + \underline{k} \cdot \underline{r} - \underline{k} \cdot \underline{v}_b t \quad (2.2)$$

where ω = wave frequency,

$\underline{k} = k_x \hat{x} + k_y \hat{y} + k_z \hat{z}$ = wavenumber vector, and

$\underline{v}_b = v_{bx} \hat{x} + v_{by} \hat{y} + v_{bz} \hat{z}$ = background atmospheric wind velocity.

The wave phase is a function of time(t), horizontal position (x, y), and altitude(z). Hence, the altitude of a Na density maximum or minimum varies as a function of time and location. By measuring the times, horizontal positions, and altitudes of a density maximum or minimum, the intrinsic wave parameters including ω and \underline{k} can be determined. Equation (2.2) can be rearranged so that the altitude variation of a density maximum or minimum can be written as a function of time and horizontal position,

$$z = a_1 t + a_2 x + a_3 y + a_4 \quad (2.3)$$

$$\text{where } a_1 = -\frac{\lambda_z}{T} + \frac{\lambda_z}{\lambda_h} v_{bh} \cos(\alpha_b - \alpha_w) + v_{bz} \quad (2.4)$$

$$a_2 = -\frac{\lambda_z}{\lambda_h} \sin \alpha_w \quad (2.5)$$

$$a_3 = -\frac{\lambda_z}{\lambda_h} \cos \alpha_w \quad (2.6)$$

$$a_4 = \frac{\theta_0}{k_z} \quad (2.7)$$

- λ_z = intrinsic vertical wavelength of the gravity wave,
- λ_h = intrinsic horizontal wavelength of the gravity wave,
- T = intrinsic wave period,
- v_{bh} = horizontal background wind velocity amplitude,
- v_{bz} = vertical background wind velocity amplitude,
- α_w = azimuth angle of the horizontal propagation direction of the gravity wave,
- α_b = azimuth angle of the horizontal propagation direction of the background wind, and
- θ_0 = wave phase at the density maximum or minimum.

The relationship between the horizontal components of the background wind and gravity wave is illustrated in Figure 2.1. As can be seen in Equations (2.3) and (2.4), the altitude variations of a Na density maximum or minimum are influenced by background wind. Equation (2.3), along with the gravity wave polarization and dispersion relations and other information which can be calculated from the sodium density profiles, can be used to determine the intrinsic parameters of the wave. For example, coefficients a_2 and a_3 defined in Equations (2.5) and (2.6) are related to the intrinsic horizontal and vertical wavelengths and the horizontal propagation direction of the wave.

In order to determine the four coefficients a_1 through a_4 uniquely, at least four measurements of a density maximum or minimum are needed. The altitude of a density maximum or minimum must be measured at an appropriate combination of four different times and horizontal positions. For example, at least three measurements at three different positions (not located in a line) and one additional measurement at one of the three positions at a different time would be sufficient. The configuration of the three horizontal positions must form a triangle to determine uniquely the wave propagation direction.

Since the wave phase is not of interest in this study and coefficient a_4 is constant, this coefficient can be eliminated from the analysis by considering the differences between the altitude measurements at three different positions. By using this approach a minimum of three

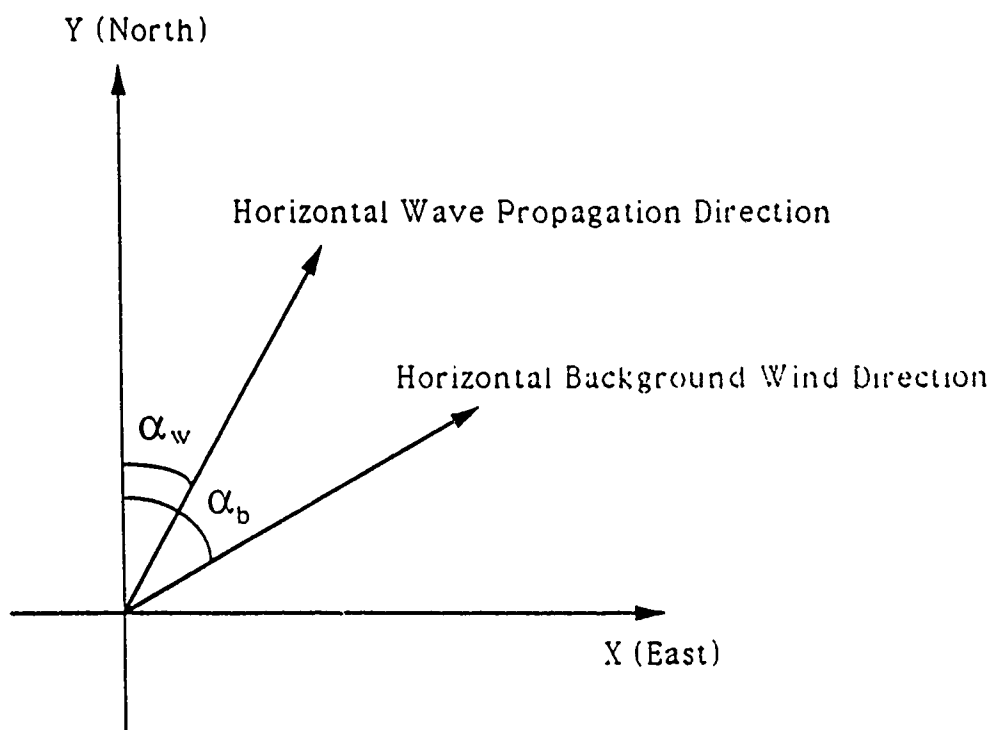


Figure 2.1. The relationship between the horizontal components of the wave propagation direction and background atmospheric wind.

measurements at three positions is sufficient to determine the coefficients a_1 through a_3 in Equation (2.3). Usually many more than three measurements will be available so that a regression analysis can be used to estimate all four coefficients, thereby reducing the error.

In order to minimize the effects of roundoff errors in computing the regression coefficients, it is the usual practise to subtract a reference data point from each measurement before the regression coefficients are calculated. Typically, the reference data point is the average of all the measurements. To simplify the mathematics, subtraction of the reference data point will not be included in the following analysis. Assume that a total of n measurements of a density maximum or minimum are obtained at several locations and times. The resulting system of equations can be written in matrix form using Equation (2.3).

$$Z = U A \quad (2.8)$$

$$\text{where } Z = [z_1 \ z_2 \ \dots \ z_n]^T \quad (2.9)$$

$$U = \begin{bmatrix} t_1 & x_1 & y_1 & 1 \\ t_2 & x_2 & y_2 & 1 \\ : & : & : & : \\ t_n & x_n & y_n & 1 \end{bmatrix} \quad (2.10)$$

$$A = [a_1 \ a_2 \ a_3 \ a_4]^T \quad (2.11)$$

where z_i , t_i , x_i , and y_i correspond to the altitude, time, x-coordinate, and y-coordinate of the i^{th} measurement. By assuming that the errors in the n measurements are mutually statistically independent, the coefficients a_1 through a_4 are computed by using the least-squares solution,

$$A = [U^T U]^{-1} U^T Z \quad (2.12)$$

The model given by Equation (2.3) assumes that there is a three-dimensional plane on which the wave phase is constant. The least-squares solution is a multiple linear regression fit of the measurements to this three-dimensional plane.

The wave propagation direction and ratio of the horizontal and vertical wavelengths can be calculated from coefficients a_2 and a_3 .

$$\alpha_w = \tan^{-1}\left(\frac{a_2}{a_3}\right) \quad (2.13)$$

$$\frac{\lambda_h}{\lambda_z} = \frac{1}{\sqrt{(a_2^2 + a_3^2)}} \quad (2.14)$$

The intrinsic vertical wavelength and wave amplitude are usually determined from the vertical wavenumber power spectra of the Na density profiles [*Gardner and Voelz, 1987*]. The vertical wavelength can also be determined by measuring the spacings between maxima and minima in the Na density profiles. The horizontal wavelength can then be calculated by multiplying Equation (2.14) by the vertical wavelength. The intrinsic wave period is determined by using the gravity wave dispersion relation,

$$T \approx \frac{\lambda_h}{\lambda_z} T_B \quad (2.15)$$

where T_B = Brunt-Vaisala period (≈ 5 min).

The measurement accuracies for these wave parameters depend on the accuracies of the computed coefficients, a_2 and a_3 , which in turn depend on the accuracies of the measured values of the altitude, z . The covariance matrix for the coefficients a_1 through a_4 estimated from mutually statistically independent observations is given by

$$C = [U^T U]^{-1} \text{Var}(z) \quad (2.16)$$

where

$$C = \begin{bmatrix} \text{Var}(a_1) & \text{Cov}(a_1, a_2) & \text{Cov}(a_1, a_3) & \text{Cov}(a_1, a_4) \\ \text{Cov}(a_1, a_2) & \text{Var}(a_2) & \text{Cov}(a_2, a_3) & \text{Cov}(a_2, a_4) \\ \text{Cov}(a_1, a_3) & \text{Cov}(a_2, a_3) & \text{Var}(a_3) & \text{Cov}(a_3, a_4) \\ \text{Cov}(a_1, a_4) & \text{Cov}(a_2, a_4) & \text{Cov}(a_3, a_4) & \text{Var}(a_4) \end{bmatrix} \quad (2.17)$$

and $\text{Var}(z)$ is the variance of the error in the altitude measurements. The variances of the computed wave parameters are

$$\text{Var}(\alpha_w) = \frac{a_3^2 \text{Var}(a_2) - 2a_2 a_3 \text{Cov}(a_2, a_3) + a_2^2 \text{Var}(a_3)}{(a_2^2 + a_3^2)^2} \quad (2.18)$$

$$\text{Var}\left(\frac{\lambda_h}{\lambda_z}\right) = \frac{a_2^2 \text{Var}(a_2) + 2a_2 a_3 \text{Cov}(a_2, a_3) + a_3^2 \text{Var}(a_3)}{(a_2^2 + a_3^2)^3} \quad (2.19)$$

$$\text{Var}(\lambda_h) = \lambda_z^2 \text{Var}\left(\frac{\lambda_h}{\lambda_z}\right) + \left(\frac{\lambda_h}{\lambda_z}\right)^2 \text{Var}(\lambda_z) \quad (2.20)$$

$$\text{Var}(T) = T_B^2 \text{Var}\left(\frac{\lambda_h}{\lambda_z}\right) + \left(\frac{\lambda_h}{\lambda_z}\right)^2 \text{Var}(T_B) \quad (2.21)$$

Unfortunately, the variance expressions given in Equations (2.18) - (2.21) do not provide any physical insights for the expected errors in the measurements. Therefore, two examples of an airborne lidar experiment and a multiple ground-based lidar experiment will be presented to illustrate the major factors influencing the measurement errors.

Consider an aircraft flying over a circular path of radius R during a total observation period of T_{ob} . The ground track of the flight is illustrated in Figure 2.2. Assume that a total of n measurements equally spaced along the flight path are obtained. The covariance matrix for this experimental configuration is calculated in Appendix 1 along with the rms errors in the gravity wave parameters. The results are

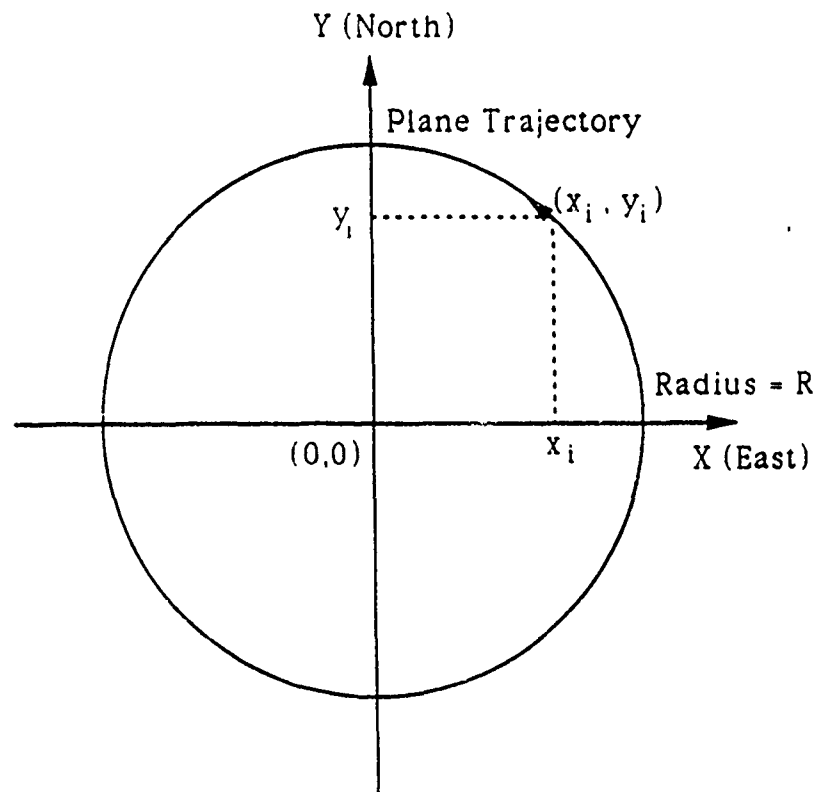


Figure 2.2. Ground track of the circular flight path with radius R .

$$\begin{aligned}\text{Std}(\alpha_w) &= \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \sqrt{\frac{\pi^2}{\pi^2 - 6} \sin^2 \alpha_w + \cos^2 \alpha_w} \\ &\approx \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}}\end{aligned}\quad (2.22)$$

$$\begin{aligned}\frac{\text{Std}\left(\frac{\lambda_h}{\lambda_z}\right)}{\left(\frac{\lambda_h}{\lambda_z}\right)} &= \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \sqrt{\sin^2 \alpha_w + \frac{\pi^2}{\pi^2 - 6} \cos^2 \alpha_w} \\ &\approx \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}}\end{aligned}\quad (2.23)$$

The rms errors in α_w and λ_h/λ_z are proportional to the rms error in the measured altitude of the density maximum or minimum, inversely proportional to the radius of the flight path, and inversely proportional to the square root of the number of measurements. The rms error in the altitude measurements is influenced by three types of noise – signal shot noise, background noise, and "geophysical noise." The shot noise is characteristic of the photon counting process used to detect the backscattered lidar signal. The background noise is introduced by the background sky light from the stars, moon and sun. Shot noise or background noise in different profiles are statistically independent. To minimize the effects of shot noise and background noise, Na density profiles are usually low-pass filtered vertically. The "geophysical noise" is usually introduced by two types of processes – background wind variations and multiple waves simultaneously influencing the Na layer. When the background wind is not constant over either the horizontal or vertical extent of the measurements, the wind could influence the variations of the altitude of a density maximum or minimum as seen in Equations (2.3) and (2.4). This effect will result in error. Multiple waves can also influence the altitude of the density perturbations and contribute to the error. For example, during the roundtrip flight from Denver to the Pacific Coast in November 1986, a dominant wave induced both a density maximum and minimum in the Na layer. By using the least-squares solution in Equation (2.12), the coefficients a_1 through

a_4 were calculated for both the density maximum and minimum. The altitudes of the density maximum and minimum were then recalculated by substituting the estimated coefficients along with the measured times and horizontal positions into Equation (2.3). The rms deviation of the difference between the measured altitudes and the least-squares model was approximately 290 m for the density maximum, and 230 m for the density minimum. Geophysical noise was the major cause of these variations. Thus typical values of $\text{Std}(z)$ will vary between 200 and 500 m. Depending on the profile integration time, geophysical noise in different profiles may be highly correlated. If so, the number of measurements, n , in Equations (2.22) and (2.23) must be set equal to the total number of statistically independent measurements.

The flight radius, R , also influences the measurement errors. The error in estimating the three-dimensional plane of a constant wave phase will be smaller, when the observations are made over a larger horizontal extent. Notice that the errors in α_w and λ_h/λ_z are also proportional to $\lambda_h/\lambda_z \approx T/T_B$. To maintain a given level of accuracy, the flight path radius should be large when observing waves with a long period or large horizontal wavelength. However, the radius can not be chosen arbitrarily large because the wave front must be coherent at each measurement point and the measurements must all be made on the same wave front. A radius comparable to the horizontal wavelength is probably a good compromise.

Consider now a configuration of three ground-based lidars at the corners of an equilateral triangle with sides of length R as illustrated in Figure 2.3. Assume that a total of n independent measurements are obtained simultaneously at each of the three lidar sites during a total observation period of T_{ob} . The covariance matrix and the measurement errors are calculated for this configuration in Appendix 2. The results are

$$\text{Std}(\alpha_w) = \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \quad (2.24)$$

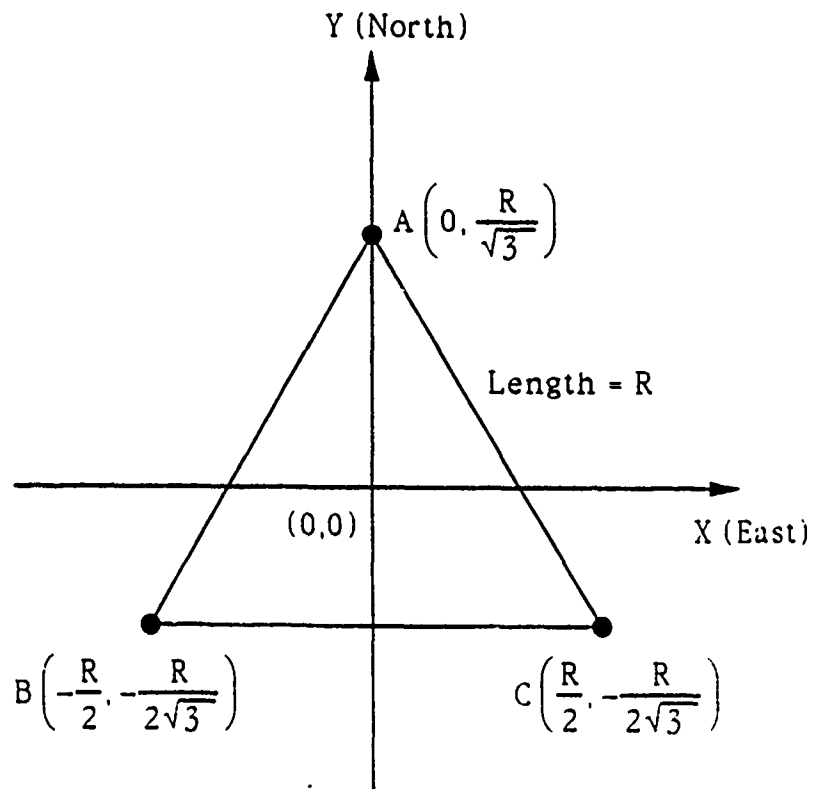


Figure 2.3. Configuration for three ground-based lidars located at the corners of an equilateral triangle with sides of length R .

$$\frac{\text{Std}\left(\frac{\lambda_h}{\lambda_z}\right)}{\left(\frac{\lambda_h}{\lambda_z}\right)} = \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \quad (2.25)$$

Notice that the equations for the measurement errors for the ground-based experiment are quite similar to those for the circular flight experiment.

The measurement error in λ_z (see [Gardner and Voelz, 1987]) and the error on the assumed value of T_B are usually negligible. A measurement accuracy of 0.1 rad ($\approx 6^\circ$) or better for the propagation direction and a measurement accuracy of 10 % or better for the intrinsic period and horizontal wavelength would require

$$\frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \leq 0.1 \quad (2.26)$$

Waves usually remain coherent in Na density profiles for 1 to 2 hours. Modern lidars can obtain accurate measurements of the Na density profile in 1 to 2 min. Thus, n is typically 50-100 under normal conditions by assuming each measurement is statistically independent of every other measurement. As mentioned previously, $\text{Std}(z)$ is dominated by geophysical noise with values typically in the range of 200 to 500 m. By assuming that $n = 50$, $\text{Std}(z) = 500$ m and $R \sim \lambda_h$, Equation (2.26) is satisfied for all waves with $\lambda_z \geq 1$ km.

In addition to the airborne and multiple ground-based measurements, steerable lidar measurements can also be used to determine the gravity wave parameters. Because of the increased atmospheric attenuation and longer propagation path lengths at lower elevation angles, the maximum zenith angle for the steerable measurements is usually limited to about 45° . At the altitudes of the Na layer near 90-100 km, a laser beam directed at the zenith angle of 45° would intersect the Na layer at a horizontal distance of about 100 km from the point directly above the lidar site. Thus, the longest practical horizontal baseline for the steerable measurements would be approximately 200 km. Based on the calculations presented above, it appears that a steerable

lidar can be used to measure accurately the intrinsic parameters of gravity waves with horizontal wavelengths of about 400 km or less.

2.3 Experimental Data

In early November of 1986, the UIUC Na lidar system was installed on the Electra aircraft operated by NCAR-RAF. Major lidar and aircraft parameters are summarized in Table 2.1. Following the installation, a total of three flights were conducted out of Stapleton International Airport in Denver, Colorado from November 13 to 18, 1986. The ground tracks of these flights are illustrated in Figure 2.4. The flight and system performance characteristics are summarized in Table 2.2. During the three flights, a total of 425 Na density profiles were collected. The integration period for each profile was 100 s. The horizontal distance corresponding to this integration period varied from 11 to 20 km, depending on the ground speed of the aircraft.

In this section, the characteristics of a quasi-monochromatic wave observed during the westward flight conducted on the night of November 17-18 will be analyzed. The flight included one eastbound leg and one westbound leg. In Figure 2.5, two sequences of Na density profiles collected during the two flight legs are plotted versus longitude in the range from 117°W to 126°W. The profiles have been filtered vertically with a cutoff of 3 km and horizontally with a cutoff of 50 km. Also the profiles have been normalized so that each has the same column abundance, and are plotted on a linear scale. At the flight altitude of approximately 8 km, the prevailing winds were eastward. Consequently, the ground speed of the aircraft was much slower during the westbound leg, and the horizontal distance corresponding to the profile integration period was much shorter. On the westbound leg in the longitude range from 113°W to 117°W, overhead clouds obscured the lidar, and no data were collected. Note the dominant density perturbations in the altitude range from 85 to 90 km in the density profiles measured on both the westbound and eastbound legs. These perturbations appear to be influenced by a quasi-monochromatic wave.

**Table 2.1. The Parameters of the UIUC Na Lidar System
and NCAR Electra Aircraft**

Transmitter

Laser	Flashlamp-pumped dye laser (Candela LFDL-1)
Wavelength	589 nm
Energy	50 mJ/pulse
Pulse Width	2 μ s FWHM
Repetition Rate	7.5 Hz
Beam Divergence	3 mrad FW @ e^{-2}

Receiver

Telescope	35 cm diameter Cassegrain (Celestron 14)
Aperture Area	0.1 m ²
Field-of-view	3 mrad
Optical Bandwidth	0.5 nm FWHM
Vertical Range Resolution	150 m

Aircraft

Model	Electra 4 engine turboprop
Nominal Cruising Altitude	6 - 8 km
Ground Speed During Observations	110 - 200 m s ⁻¹

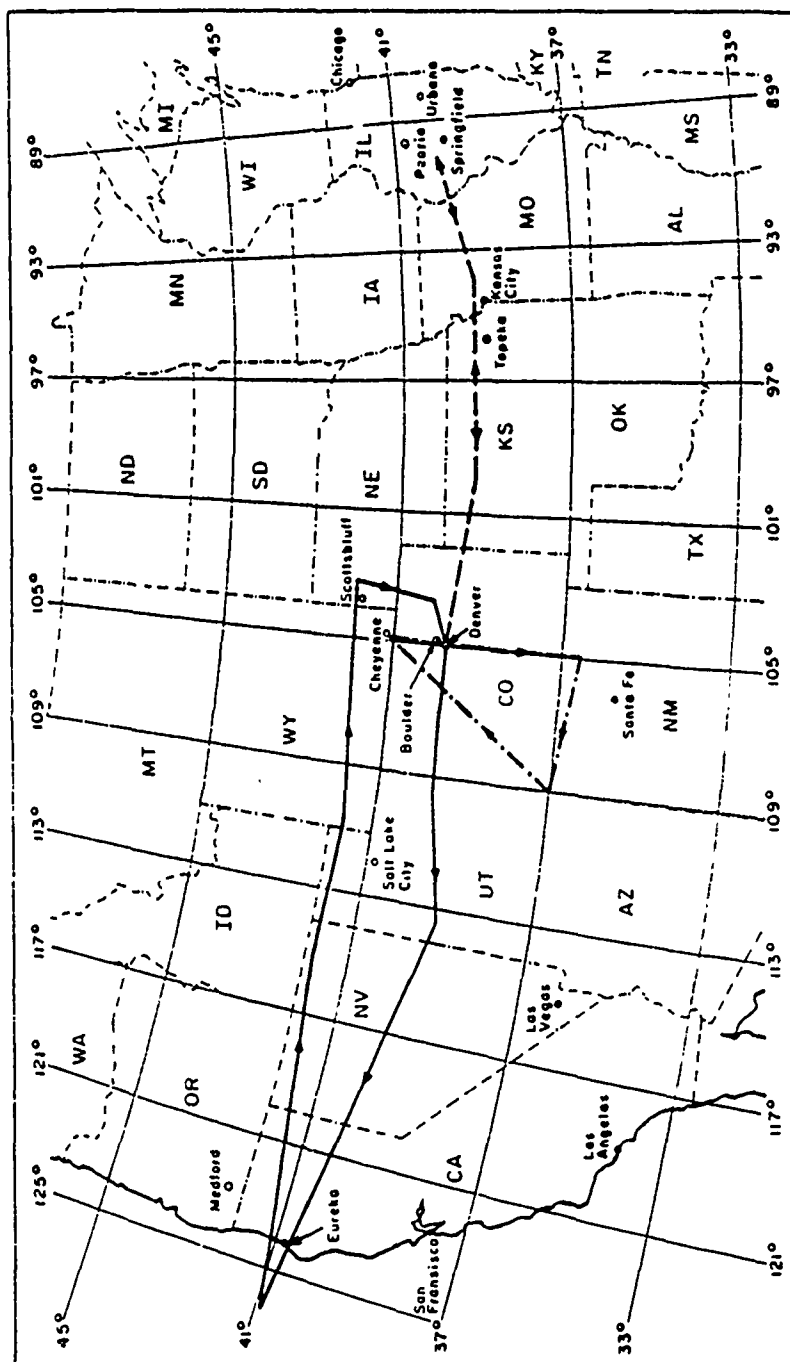


Figure 2.4. Ground tracks of the three flights of the airborne Na lidar experiment in November, 1986. The flights were conducted out of Denver, Colorado.

Table 2.2. Summary of the Flights and System Performance

Flight	1	2	3
Date	Nov. 13	Nov. 15-16	Nov. 17-18
Observation Time (MST)	2119-2357	2210-0309	2142-0502
Observation Duration (hours)	2.6	5.0	7.3
Flight Pattern	Triangular	Eastward	Westward
Location	CO,NM,AZ,WY	CO-IL	CO-Pacific Coast
Na Signal Level ^a (total Na counts/shot)	10	7	8

Total Observation Time = 14.9 hours

^aTypical ground-based signal level at Urbana is 5 counts/shot.

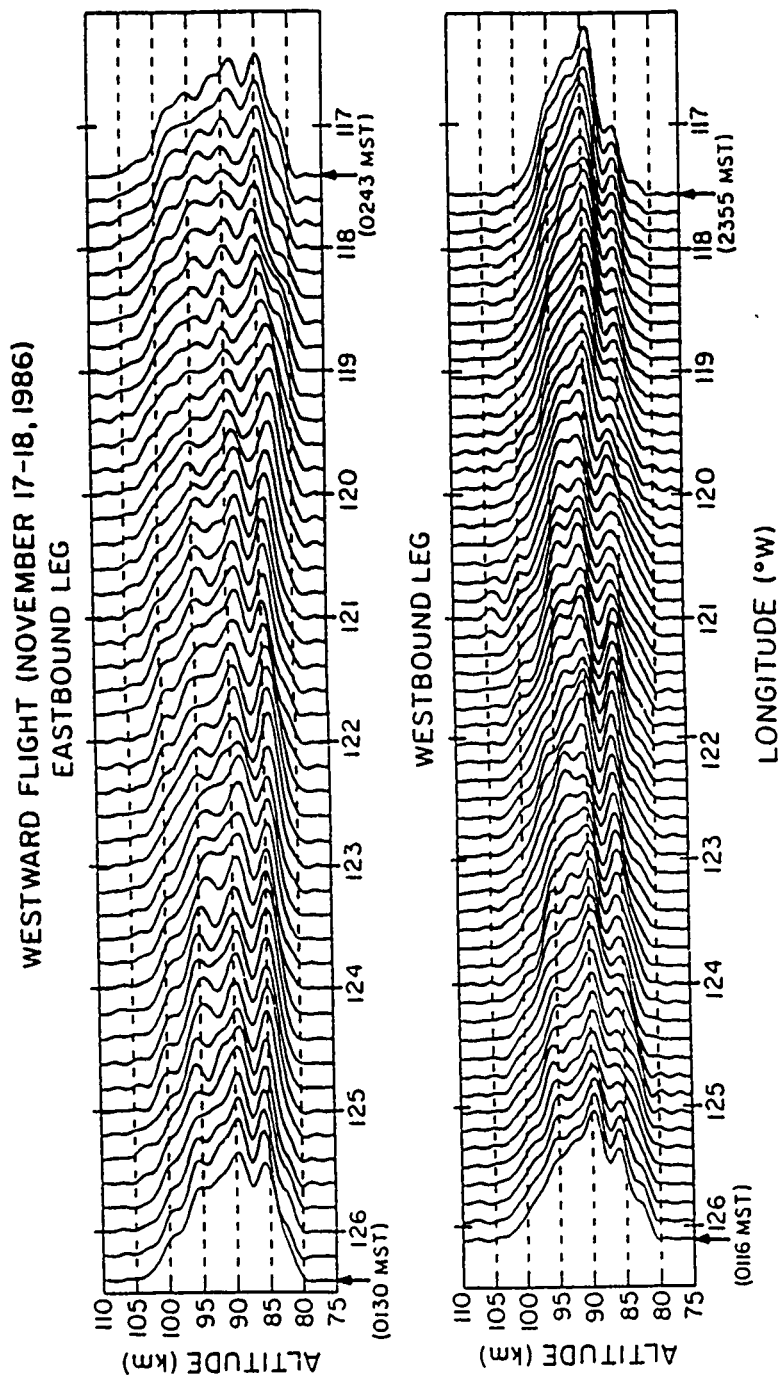


Figure 2.5. Sodium density profiles collected during the westbound and eastbound legs of the westward flight on November 17-18, 1986. The profiles have been filtered vertically with a cutoff of 3 km and horizontally with a cutoff of 50 km. The profiles have been also normalized so that each has the same column abundance, and are plotted on a linear scale.

The altitude variations of a local Na density maximum and minimum induced by the dominant wave on the bottomside of the layer are plotted versus time in Figure 2.6. The altitudes were computed by first filtering the Na density profiles vertically with a cutoff of 2 km. Between 0115 and 0130 MST, the aircraft changed flight direction from approximately 277° azimuth (westward) to 80° azimuth (eastward), and no profiles were collected during this period. Circles represent the altitudes of the density maximum, and crosses represent the altitudes of the density minimum. A total of 46 measurements of the density maximum were obtained. The averages of the measured times, horizontal positions, and altitudes were computed and used for the reference data point in the regression model. The coefficients a_1 through a_4 were calculated by using the least-squares solution in Equation (2.12), and the altitudes of the density maximum were then recalculated by substituting the estimated coefficients along with the measured times and positions into Equation (2.3). The results are plotted as the solid line near the altitudes of the density maximum in Figure 2.6. The standard deviation of the measured altitudes from the least-squares fitted altitudes was 290 m. The data for the density minimum are processed in a similar manner. A total of 50 measurements were obtained, and the solid line near the altitudes of the density minimum is the least-squares fit for these measurements. The standard deviation of the measured altitudes from the least-squares fitted altitudes was 230 m.

The vertical wavelength of the wave can be estimated using several approaches. The average vertical separation between the least-squares fits of the density maxima and minima plotted in Figure 2.6 is approximately equal to $\lambda_z/2$. However, the steep density gradients on the bottomside of the Na layer will result in separation distances that are slightly shorter than $\lambda_z/2$. For the data plotted in Figure 2.6, the least-squares fits are separated by an average distance of 1.9 km so that λ_z is slightly larger than 3.8 km.

The vertical wavelength of the dominant wave can also be estimated from the vertical wavenumber power spectrum of the Na profiles. The technique is described in detail by *Gardner and Voelz* [1987]. The power spectrum of the Na profiles is plotted in Figure 2.7. The vertical wavelength of the dominant wave which was responsible for the density perturbations in the

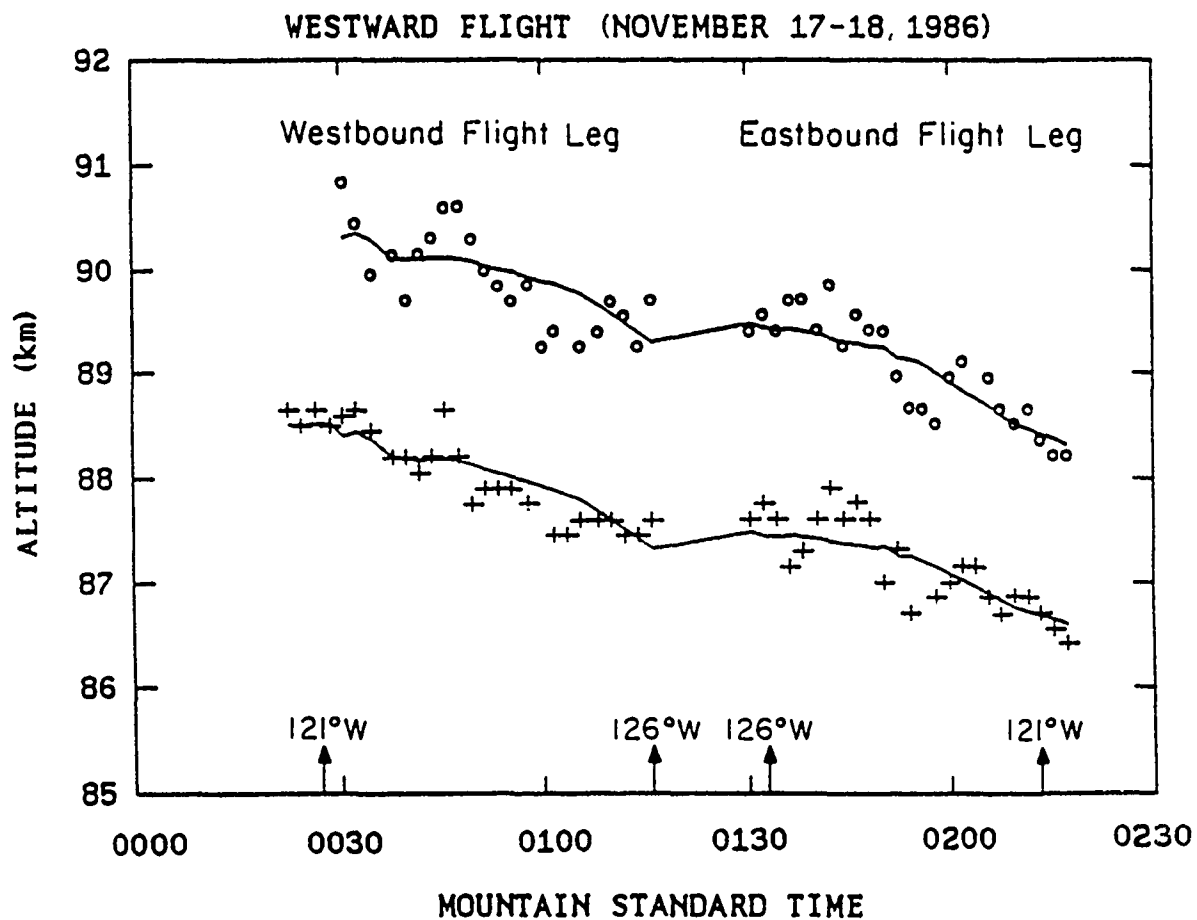


Figure 2.6. The altitude variations of a local Na density maximum and minimum measured during the westward flight on November 17-18, 1986. Circles represent the altitudes of the density maximum, and crosses represent the altitudes of the density minimum. The solid lines represent the least-squares fitted altitudes of the density maximum and minimum.

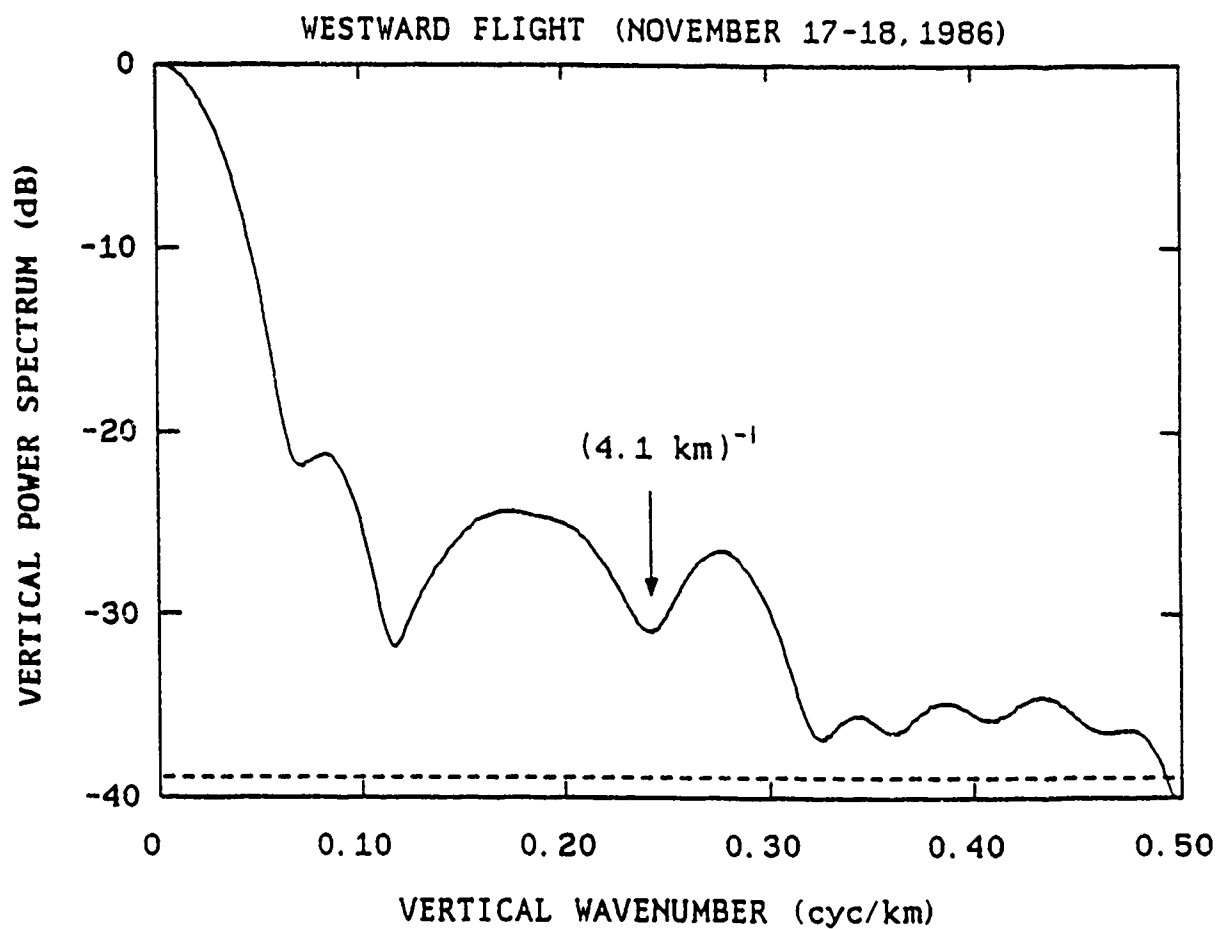


Figure 2.7. Average vertical wavenumber power spectrum of the Na density profiles collected from 0131 to 0138 MST during the westward flight on November 17-18, 1986.

altitude region from 85 to 90 km (Figure 2.5) is believed to be 4.1 km. The amplitude of this wave is also calculated from the spectrum to be 16 m s^{-1} [Gardner and Voelz, 1987].

The coefficients estimated from the least-squares solutions were used in Equations (2.13) through (2.15) to calculate the gravity wave propagation direction, ratio of horizontal to vertical wavelength, and intrinsic period. The vertical wavelength, wave amplitude, and amplitude growth length were calculated from the Na profile power spectrum. The results are summarized in Table 2.3. The wave parameters calculated from the variations of the density maximum are quite comparable to those calculated from the variations of the density minimum, which indicates that the dominant wave was influencing both the density maximum and minimum. The wave propagation angle was almost due south. The large values for the standard deviations of the estimated horizontal wavelengths and periods were the results of large variances computed for the ratio of the horizontal to vertical wavelengths, λ_x/λ_z , in Equation (2.19). The large variances appear to be the result of the directions of the two flight legs which were both approximately orthogonal (westbound and eastbound) to the wave propagation direction (southbound). By assuming the $v_{bz} \approx 0$, it is also possible to estimate the background horizontal wind velocity in the direction of wave propagation by using Equation (2.4). The calculated background horizontal wind velocity was about 5 m s^{-1} northward, which is comparable to the meridional background wind measured at Platteville, Colorado (40°N , 105°W) near 90 km altitude during the period from November 4 - 20, 1986 [Kwon *et al.*, 1989b]. The horizontal distance between the location of the airborne lidar observations and the location of the source of a wave can be estimated roughly by multiplying the height of the Na layer by the ratio of the intrinsic horizontal wavelength to the vertical wavelength of the observed wave. For the dominant wave observed during the flight, the horizontal distance between the measurement locations and the tropospheric source was computed to be approximately 1800 km. Interestingly, during the period from November 15 to 18, a meteorological low pressure system developed over the Gulf of Alaska, which is about 2800 km north of the flight path. Although it is speculative, this low pressure system might be responsible for the observed wave.

**Table 2.3. Intrinsic Parameters of the Dominant Wave Observed
During the Westward Flight on November 17-18, 1986**

Altitude Variations	Density Maximum	Density Minimum
Data Points	46	50
Vertical Wavelength (km)	4.1 (± 0.03)	4.1 (± 0.03)
Horizontal Wavelength (km)	80 (± 28)	89 (± 27)
Period(min)	97 ($\pm 34^a$)	107 ($\pm 32^a$)
Wave Propagation Direction ($^\circ$)	178.1 (± 0.4)	178.6 (± 0.3)
Wave Amplitude ($m s^{-1}$)	17 (± 1)	17 (± 1)
Amplitude growth length (km)	16 (± 2)	16 (± 2)

^aCalculated by assuming $Var(T_B) \approx 0$ in Equation (2.21).

There are other waves influencing the Na layer during the flight which appear to be propagating westward. The horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound and eastbound legs of the westward flight are plotted in Figures 2.8 a and b, respectively. The technique for estimating these horizontal wavenumber spectra is described in detail by *Kwon et al.* [1989a]. The horizontal wavenumber is the inverse of zonal distance along the flight legs. Note the spectral peaks in the bottomside spectrum for the westbound observations (Figure 2.8 a) near the wavelengths of 457 km and 316 km, and the peaks in the bottomside spectrum for the eastbound observations (Figure 2.8 b) near 319 km and 207 km. The peak near 457 km in Figure 2.8 a appears to correspond to the peak near 319 km in Figure 2.8 b, and the peak near 316 km in Figure 2.8 a appears to correspond to the peak near 207 km in Figure 2.8 b. These peaks may be related to westward propagating waves. In the bottomside spectrum measured on the westbound observations (Figure 2.8 a), these peaks seem to be Doppler shifted to lower wavenumbers. The observed Doppler shifted wavelength and the intrinsic zonal wavelength are related by the following equation [*Kwon et al.*, 1989a].

$$\lambda_{in} = \lambda_{ob} \left(1 - \frac{v_{ob}}{v_a} \right) \quad (2.27)$$

where λ_{in} = intrinsic wavelength along the flight path,

λ_{ob} = observed wavelength along the flight path,

$v_{ob} = v_p + v_b$ = observed phase velocity along the flight path,

v_p = intrinsic phase velocity along the flight path,

v_b = background atmospheric wind velocity along the flight path at the altitudes of the Na layer, and

v_a = aircraft velocity.

Because observations were made over both the westbound and eastbound flight legs, and the aircraft velocity is known, Equation (2.27) can be solved for both the intrinsic wavelength and observed phase velocity along the flight path. By using the two observed wavelengths of

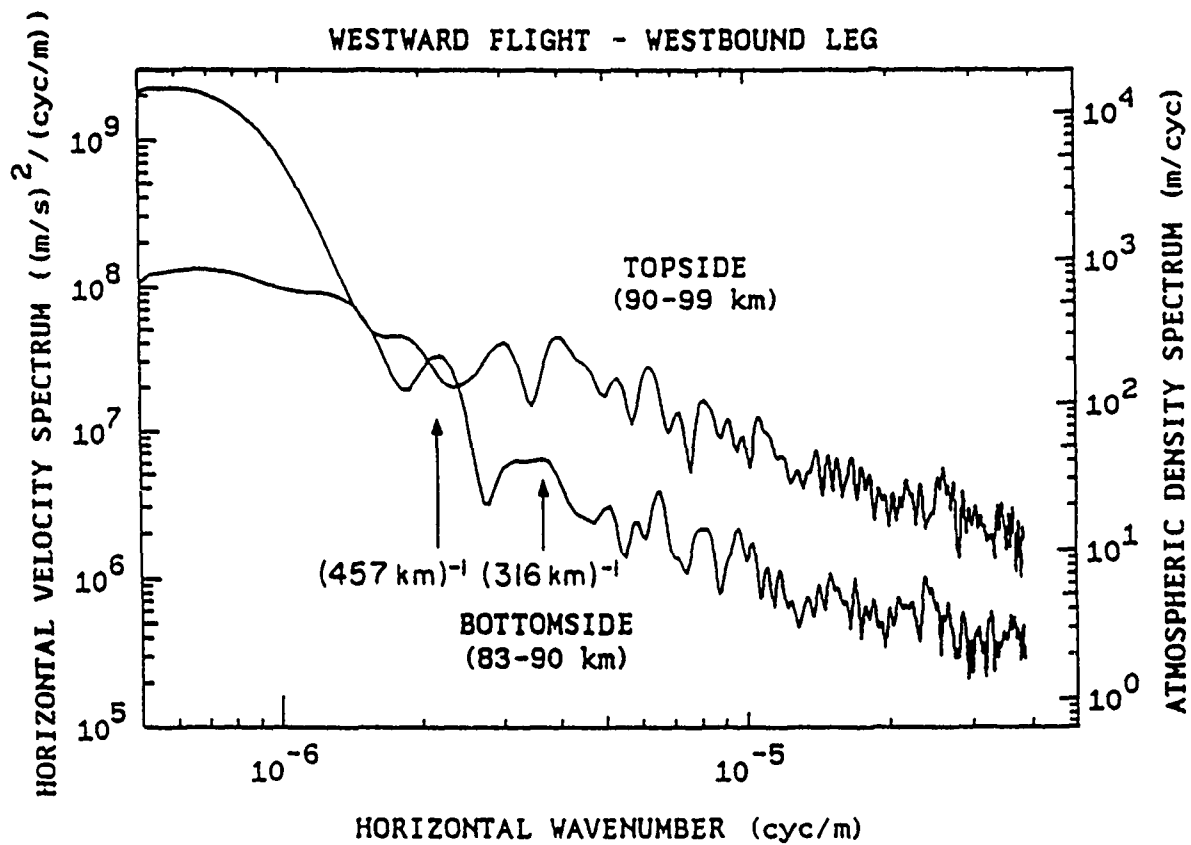


Figure 2.8 a). Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound leg of the westward flight on November 17-18, 1986. The data were filtered vertically with a cutoff of 1 km before the spectra were computed.

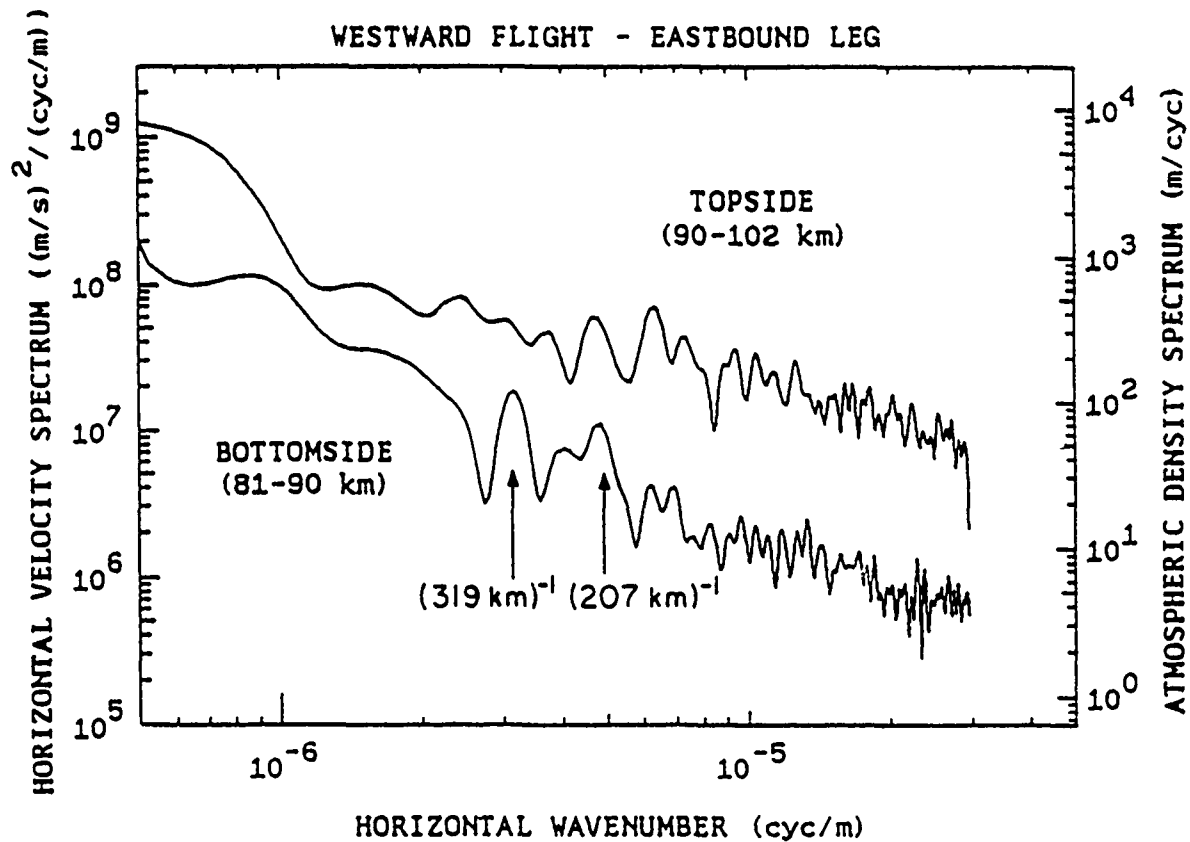


Figure 2.8 b). Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the eastbound leg of the westward flight on November 17-18, 1986. The data were filtered vertically with a cutoff of 1 km before the spectra were computed.

316 km and 207 km, the intrinsic zonal wavelength of this wave is computed to be 243 km, and the observed phase velocity is approximately 31 m s^{-1} westward. The average zonal wind velocity measured at Platteville, Colorado was about 5 m s^{-1} eastward over the altitude range of the Na layer [Kwon *et al.*, 1989b]. By subtracting this background wind velocity from the observed phase velocity, the intrinsic zonal phase velocity is calculated to be 36 m s^{-1} westward. The intrinsic wave period is then calculated to be 111 min, or approximately 2 hours. Therefore, there appeared to be at least two waves with the periods of 102 min and 111 min influencing the Na layer during the flight; one wave was propagating westward, and the other wave was propagating southward. For the wave corresponding to the spectral peak at 457 km in Figure 2.8 a, and the peak at 319 km in Figure 2.8 b, the intrinsic zonal wavelength is approximately 366 km, the intrinsic zonal phase velocity is 32 m s^{-1} westward, and the intrinsic period is 192 min.

Dominant waves with the period of approximately 2 hours have been observed often with the UTUC Na lidars at Urbana, Illinois (40°N) [Gardner *et al.*, 1986; Gardner and Voelz, 1987; Gardner, 1989], at Broomfield, Colorado (40°N) [Kwon *et al.*, 1989b], and during the eastward flight on November 15-16, 1986 which was conducted over the Great Plains in the latitude range from 39°N to 40°N (Figure 2.4) [Kwon *et al.*, 1989a]. These waves were dominant only on the bottomside of the Na layer (80-90 km). The zonal component of the 2-hour period wave measured during the eastward flight was observed to propagate westward at an apparent velocity of 38 m s^{-1} , and the intrinsic zonal wavelength of this wave was 263 km.

In order to estimate spectral slopes for the horizontal wavenumber spectra plotted in Figures 2.8 a and b, linear regression fits were performed over horizontal scales from 70 to 700 km. The spectral slopes for the bottomside and topside spectra for the westbound observations (Figure 2.8 a) were -1.69 and -1.05, respectively. The slopes for the bottomside and topside spectra for the eastbound observations (Figure 2.8 b) were -1.41 and -0.71, respectively. The spectral slopes for the topside spectra were consistently shallower than those of the bottomside spectra. The shallower slopes of the topside spectra may be the result of saturation affecting the

larger scale waves. The topside spectral amplitudes were larger which indicates that the wave amplitudes were growing with altitude.

2.4 Summary

A data analysis technique for determining intrinsic gravity wave parameters including wave propagation direction has been presented. This technique involves measuring the altitude variations of density perturbations induced in the mesospheric Na layer by gravity waves. This technique can be used with airborne lidars, multiple ground-based lidars, and steerable lidars. Several examples have been presented to show the expected measurement accuracies for the wave parameters. Although the technique was demonstrated using Na lidar data, it can also be used with Rayleigh lidars.

The technique has been successfully applied to the airborne Na lidar data obtained during the roundtrip flight from Denver, Colorado to the Pacific Coast in November 1986. The horizontal wavelength of a dominant wave observed during the flight was approximately 85 km, and the vertical wavelength was 4.1 km. The intrinsic period was 102 min, and the propagation direction was almost due south.

The wave parameters of two additional gravity waves have been estimated from the horizontal wavenumber spectra measured during the flight. One of the waves had an approximate period of 2 hours, and its zonal component was propagating westward at a velocity of 31 m s^{-1} . The parameters of this wave were quite similar to those of the 2-hour period wave observed during the eastward flight on November 15-16, 1986. Based upon these measurements and ground-based observations made at Urbana, Illinois, it appears that the 2-hour period waves often dominate the bottomside of the Na layer (80-90 km) at mid-latitudes over North America.

3. AIRBORNE SODIUM LIDAR OBSERVATIONS OF HORIZONTAL AND VERTICAL WAVENUMBER SPECTRA OF MESOPAUSE DENSITY AND WIND PERTURBATIONS

3.1 Introduction

Gravity waves play an important role in the dynamics of the mesopause region. Because this region is relatively inaccessible to in-situ measurements, most studies have relied on remote sensing techniques. In recent years, Na lidar techniques have been developed to provide important data on gravity waves and tides near the mesopause. The Na layer is generally confined to an altitude range from 80 to 110 km, and the peak density typically occurs at 90 km. The Na layer is particularly useful for studying mesopause dynamics, because the layer profile is very sensitive to gravity wave perturbations. The amplitudes of the waves are usually large near the mesopause, and the steep Na density gradients on the bottom and topsides of the layer tend to enhance the observed wave perturbations. *Gardner and Voelz* [1987] reported an extensive analysis of the monochromatic gravity waves observed in the Na layer above Urbana, Illinois. Studies of atmospheric tides using Na lidars have also been reported by *Batista et al.* [1985] and *Kwon et al.* [1987].

Because most radar and lidar observations have been made at fixed elevation angles, studies of gravity waves are usually limited to exploring only the vertical and temporal structures of the waves. Radar-based studies of gravity waves have shown an approximate k_z^{-3} dependence in vertical wavenumber spectra of the horizontal wind perturbations [*Smith et al.*, 1985; *Tsuda et al.*, 1988], and an approximate $f^{-5/3}$ dependence in the temporal frequency spectra [*Balsley and Carter*, 1982; *Vincent*, 1984]. *Fritts et al.* [1989] have presented the only measurements of the horizontal wavenumber spectra of the mesopause region. They calculated the spectra in the altitude range from 60 to 90 km from atmospheric density variations measured during seven re-entries of NASA space shuttles. The horizontal wavenumber spectra exhibited a slope of approximately -2 at horizontal scales from 10 to 1000 km.

With the advent of airborne Na lidar techniques, the study of the horizontal structure of gravity waves has become more feasible. In March 1983, the UTUC group conducted the first airborne Na lidar observations on a single roundtrip flight from the NASA Wallops Flight Facility to Albany, New York [Segal *et al.*, 1984]. This experiment demonstrated the feasibility of airborne measurements and revealed evidence of wave induced horizontal structures over the 650 km flight path.

In November of 1986, the UTUC group conducted another airborne Na lidar campaign with the support of the NCAR-RAF in Broomfield, Colorado. In order to investigate the longitudinal characteristics of gravity waves at mid-latitudes, a total of three flights were made over the Great Plains, Rocky Mountains, and Pacific Coast. The flights were conducted out of Stapleton International Airport in Denver, Colorado using the NCAR Electra aircraft. The kinetic energy horizontal and vertical wavenumber spectra of horizontal winds inferred from the airborne Na density profiles are presented in this chapter. The theoretical basis for estimating the horizontal wind amplitudes and the spectra from the Na lidar data is described in Section 3.2. The experiment is described in Section 3.3, and the experimental data are presented and discussed in Section 3.4. The airborne data are compared with the similar observations obtained with other techniques in Section 3.5.

3.2 Layer Density Response

The density response of an atmospheric layer composed of a minor neutral constituent to a wave induced wind perturbation is governed by the continuity equation. By neglecting diffusion and chemical effects and assuming the unperturbed layer is horizontally homogeneous, Gardner and Shelton [1985] have shown that the density response can be written in the form

$$n_s(\mathbf{p},t) = e^{-\phi} n_0(z - \theta_z) \quad (3.1)$$

where

$$n_s(\mathbf{p},t) = \text{Na density at position } \mathbf{p} \text{ and time } t$$

$n_0(z)$ = steady-state Na density profile in the absence of wind

perturbations

$\underline{p} = x\hat{x} + z\hat{z}$ = position vector where x is the horizontal coordinate and z is the vertical coordinate

and ϕ and θ_z are solutions to the partial differential equations

$$\frac{\partial \phi}{\partial t} = \nabla \cdot \underline{V} - \underline{V} \cdot \nabla \phi \quad (3.2)$$

$$\frac{\partial \theta_z}{\partial t} = v_z - \underline{V} \cdot \nabla \theta_z \quad (3.3)$$

where $\underline{V} = v_x \hat{x} + v_z \hat{z}$ is the atmospheric velocity field. If the wind perturbations are small the second terms on the right-hand sides of (3.2) and (3.3) can be neglected [*Gardner and Shelton, 1985*] so that

$$\phi(\underline{r}, t) \approx \int_{-\infty}^t \nabla \cdot \underline{V} \, d\tau \quad (3.4)$$

$$\theta_z(\underline{r}, t) \approx \int_{-\infty}^t v_z \, d\tau \quad (3.5)$$

In this case the layer response is given by

$$n_s(\underline{r}, t) \approx \exp \left(- \int_{-\infty}^t \nabla \cdot \underline{V} \, d\tau \right) n_0 \left(z - \int_{-\infty}^t v_z \, d\tau \right) \quad (3.6)$$

The layer perturbations result from a multiplicative distortion due to the wind divergence and a vertical displacement of the layer caused by the vertical winds. The relative importance of these two effects depends on the spatial scale of the wave causing the wind fluctuations.

For the following analysis it is most convenient to work with the natural logarithm of the ratio of the perturbed and unperturbed Na density profiles. If the unperturbed Na density profile is modeled as

$$n_0(z) = \exp [- g(z)] \quad (3.7)$$

then Equation (3.1) can be used to show that

$$r_s(p,t) = \ln(n_s/n_0) = -\phi - g(z - z_0) + g(z) \quad (3.8)$$

By expanding $g(z)$ in a Taylor series about the centroid height (z_0) of the unperturbed layer and by retaining terms out to first order in \underline{y} , $r_s(p,t)$ in Equation (3.8) can be written

$$r_s(p,t) = -\phi - g'(z) \theta_z \quad (3.9)$$

By assuming that $g(z)$ is approximately quadratic or equivalently that the unperturbed layer is approximately Gaussian, Equation (3.9) can be simplified

$$r_s(p,t) \approx -\phi + \frac{(z - z_0)}{\sigma_0^2} \theta_z \quad (3.10)$$

where σ_0 is the rms thickness of the unperturbed sodium layer. By assuming that the atmospheric density profile is approximately exponential

$$n_a(z) = n_0 e^{-(z - z_0)/H} \quad (3.11)$$

where H is the atmospheric scale height and n_0 is the density at altitude z_0 , the log-ratio of the perturbed and unperturbed atmospheric densities is [Gardner et al., 1989]

$$r_a(p,t) = -\phi + \theta_z/H \quad (3.12)$$

For gravity wave perturbations, it is possible to simplify Equations (3.10) and (3.12) for the relative density perturbations. When the vertical wavelength λ_z and frequency ω satisfy the conditions $\lambda_z \ll 4\pi H$ and $\omega \ll N$, where N is the Brunt-Vaisala frequency, the gravity wave polarization and dispersion relations can be used to show that

$$\theta_z = \gamma H \phi \quad (3.13)$$

where γ is the ratio of specific heats. In this case, the relative density perturbations can be written as

$$r_s(p,t) = -[1 - \gamma H(z - z_0)/\sigma_0^2] \phi(p,t) \quad (3.14)$$

$$r_a(p,t) = -(1 - \gamma) \phi(p,t) \quad (3.15)$$

The gravity-wave polarization and dispersion relations can be used to relate the vertical

wavenumber (k_z) spectrum, horizontal wavenumber (k_x) spectrum and mean-square value of ϕ to the vertical and horizontal wavenumber gravity wave power spectra ($E_x(k_z)$ and $E_x(k_x)$) and mean-square value of the horizontal winds ($\langle v_x^2 \rangle$) [Gardner et al, 1989].

$$E_x(k_z) = (\gamma H N)^2 E_\phi(k_z) \quad (3.16)$$

$$E_x(k_x) = (\gamma H N)^2 E_\phi(k_x) \quad (3.17)$$

$$\langle v_x^2 \rangle = (\gamma H N)^2 \langle \phi^2 \rangle \quad (3.18)$$

By combining Equations (3.14), (3.15) and (3.18) the mean-square relative sodium density perturbations can be expressed in terms of the mean-square atmospheric density perturbations or the mean-square horizontal winds.

$$\begin{aligned} \langle r_s^2(z, t) \rangle &= [1 - \gamma H(z - z_0)/\sigma_0^2]^2 \langle r_a^2(z, t) \rangle / (\gamma - 1)^2 \\ &= [1 - \gamma H(z - z_0)/\sigma_0^2]^2 \langle v_x^2(z, t) \rangle / (\gamma H N)^2 \end{aligned} \quad (3.19)$$

Because the ensemble average $\langle r_s^2 \rangle$ can be approximated by spatially averaging the relative Na density perturbations throughout the vertical extent of the layer, it is easy to show that

$$\langle v_x^2(x, t) \rangle \approx \frac{\mu^2}{\Delta z} \int_{z_0 - \Delta z/2}^{z_0 + \Delta z/2} r_s^2(z, t) dz \quad (3.20)$$

and

$$\langle r_a^2(x, t) \rangle \approx \left(\frac{\gamma - 1}{\gamma H N} \right)^2 \frac{\mu^2}{\Delta z} \int_{z_0 - \Delta z/2}^{z_0 + \Delta z/2} r_s^2(z, t) dz \quad (3.21)$$

where

$$\mu^2 = \frac{(\gamma H N)^2}{[1 + (\gamma H \Delta z / \sigma_0^2)^2 / 12]} \quad (3.22)$$

and Δz is the vertical extent of the Na layer.

The horizontal wavenumber spectra can also be related using Equations (3.14), (3.15) and (3.17). If the Na density spectra are computed at each altitude throughout the layer and then averaged, the horizontal wind spectrum, E_x , and the atmospheric density spectrum, E_a , can be shown as

$$E_x(k_x) = \frac{\mu^2}{\Delta z} \int_{z_0 - \Delta z/2}^{z_0 + \Delta z/2} \frac{\langle |R_s(k_x, z, t)|^2 \rangle}{\Delta x} dz \quad (3.23)$$

and

$$E_a(k_x) = \left(\frac{\gamma - 1}{\gamma H N} \right)^2 \frac{\mu^2}{\Delta z} \int_{z_0 - \Delta z/2}^{z_0 + \Delta z/2} \frac{\langle |R_s(k_x, z, t)|^2 \rangle}{\Delta x} dz \quad (3.24)$$

where R_s is Fourier transform of r_s

$$R_s(k_x, z, t) = \int_{x_0 - \Delta x/2}^{x_0 + \Delta x/2} r_s(x, z, t) e^{ik_x x} dx \quad (3.25)$$

and Δx is the horizontal extent of the Na observations. The horizontal wind spectrum and atmospheric density spectrum can be also calculated for a specific altitude range, i.e., the layer bottomside (from the bottomedge to the centroid), or the layer topside (from the centroid to the topedge of the layer). The bottomside horizontal wind spectrum, E_{bx} , and the bottomside atmospheric density spectrum, E_{ba} , are

$$E_{bx}(k_x) = 2 \frac{\mu_b^2}{\Delta z} \int_{z_0 - \Delta z/2}^{z_0} \frac{\langle |R_s(k_x, z, t)|^2 \rangle}{\Delta x} dz \quad (3.26)$$

$$E_{ba}(k_x) = 2 \left(\frac{\gamma - 1}{\gamma H N} \right)^2 \frac{\mu_b^2}{\Delta z} \int_{z_0 - \Delta z/2}^{z_0} \frac{\langle |R_s(k_x, z, t)|^2 \rangle}{\Delta x} dz \quad (3.27)$$

where

$$\mu_b^2 = \frac{(\gamma H N)^2}{[1 + \gamma H \Delta z / 2 \sigma_0^2 + (\gamma H \Delta z / \sigma_0^2)^2 / 12]} \quad (3.28)$$

The topside horizontal wind spectrum, E_{α} , and the topside atmospheric density spectrum, E_{α} , are

$$E_{\alpha}(k_x) \approx 2 \frac{\mu_t^2}{\Delta z} \int_{z_0}^{z_0+\Delta z/2} \frac{\langle |R_s(k_x, z, t)|^2 \rangle}{\Delta x} dz \quad (3.29)$$

$$E_{\alpha}(k_x) \approx 2 \left(\frac{\gamma-1}{\gamma H N} \right)^2 \frac{\mu_t^2}{\Delta z} \int_{z_0}^{z_0+\Delta z/2} \frac{\langle |R_s(k_x, z, t)|^2 \rangle}{\Delta x} dz \quad (3.30)$$

where

$$\mu_t^2 = \frac{(\gamma H N)^2}{[1 - \gamma H \Delta z / 2 \sigma_0^2 + (\gamma H \Delta z / \sigma_0^2)^2 / 12]} \quad (3.31)$$

If the observation altitude range Δz is large compared to the vertical correlation length of the waves, then the vertical wavenumber spectra are related by the following equations [Gardner and Senft, 1989].

$$E_x(k_z) \approx \frac{\mu^2}{\Delta x} \int_{x_0-\Delta x/2}^{x_0+\Delta x/2} \frac{\langle |R_s(x, k_z, t)|^2 \rangle}{\Delta z} dx \quad (3.32)$$

$$E_a(k_z) \approx \left(\frac{\gamma-1}{\gamma H N} \right)^2 \frac{\mu^2}{\Delta x} \int_{x_0-\Delta x/2}^{x_0+\Delta x/2} \frac{\langle |R_s(x, k_z, t)|^2 \rangle}{\Delta z} dx \quad (3.33)$$

where

$$R_s(x, k_z, t) = \int_{z_0-\Delta x/2}^{z_0+\Delta x/2} r_s(x, z, t) e^{ik_z z} dz \quad (3.34)$$

Quasi-monochromatic gravity waves are frequently observed in the Na density profiles. For the case of low-frequency monochromatic waves, exact solutions of Equations (3.2) and (3.3) for ϕ and θ_z were derived by Gardner and Shelton [1985].

$$\phi = \ln \left[1 + \frac{Ae^{\beta z}}{\gamma - 1} \cos(\omega t - \mathbf{k} \cdot \mathbf{p}) \right] \quad (3.35)$$

$$\theta_z = \ln \left[1 + \frac{Ae^{\beta z}}{\gamma - 1} \cos(\omega t - \mathbf{k} \cdot \mathbf{p}) \right] \quad (3.36)$$

where $Ae^{\beta z}$ = wave amplitude
 β = amplitude growth factor (m^{-1})
 $\mathbf{k} = k_x \hat{x} + k_z \hat{z}$ = wavenumber vector (m^{-1})
 k_x = horizontal wavenumber (m^{-1})
 k_z = vertical wavenumber (m^{-1})
 ω = wave frequency (s^{-1}).

The corresponding vertical and horizontal winds generated by an unsaturated gravity wave are given by [Hines, 1960]

$$v_z \approx - \frac{\gamma H N}{\gamma - 1} \frac{\lambda_z}{\lambda_x} Ae^{\beta z} \sin(\omega t - \mathbf{k} \cdot \mathbf{p}) \quad (3.37)$$

$$v_x \approx - \frac{\gamma H N}{\gamma - 1} \frac{\lambda_z}{\lambda_x} Ae^{\beta z} \sin(\omega t - \mathbf{k} \cdot \mathbf{p}) \quad (3.38)$$

The vertical and horizontal wavelengths are, respectively, λ_z and λ_x . By substituting Equation (3.35) into (3.14) and evaluating the horizontal wavenumber spectrum given by Equation (3.23) at the horizontal wavenumber of the wave, it can be shown that

$$E_x(k_x) \approx \frac{\Delta x}{2} v_x^2(z_0) \left[\cosh(\beta \Delta z) - \frac{2\sigma_0^2}{\gamma H \Delta z} \sinh(\beta \Delta z) \right] \quad (3.39)$$

Equation (3.39) can be used to calculate the horizontal wind velocity of monochromatic gravity waves by measuring the amplitudes of the spectral peaks in the horizontal wavenumber spectrum.

3.3 Description of the Experiment

In early November of 1986, the UIUC lidar system was installed on the Electra aircraft operated by the NCAR-RAF. The system included a flashlamp-pumped dye laser, a 35 cm diameter Cassegrain telescope, and associated electronic and optical subsystems. Figure 3.1 is a photograph of the NCAR Electra, and Figure 3.2 is a photograph of the Electra interior showing the lidar installation. The major lidar and aircraft parameters were summarized in Table 2.1. Following the installation, a total of three flights were conducted out of Stapleton International Airport in Denver from November 13 to 18, 1986. The ground tracks of these flights were illustrated in Figure 2.4. The flights and system performance were summarized in Table 2.2. During the three flights, a total of 425 Na density profiles were collected. The integration period for each profile was 100 s. The horizontal distance corresponding to this integration period ranged from 11 to 20 km, depending on the ground speed of the aircraft.

The aircraft position and altitude were monitored and recorded during the flights. The aircraft longitude and latitude were measured with an Inertial Navigation System (INS), and the altitude was measured with a pressure transducer. The INS was also used to measure the roll and pitch angles of the aircraft, and flight directions. The resolution of the latitude and longitude measurements was 0.0014° , and the resolution of the roll and pitch angles, and flight direction was 0.0028° . Radiosonde data were also collected at 5 different locations near the flight tracks; Peoria, Illinois (41°N , 90°W), Topeka, Kansas (39°N , 96°W), Denver, Colorado (40°N , 105°W), Salt Lake City, Utah (41°N , 112°W), and Medford, Oregon (42°N , 123°W). The radiosonde data revealed that the error of the measured pressure altitude due to meteorological variations was negligible during the flights. The roll angle of the aircraft was usually less than 1° , and the pitch angle was relatively constant at about 3° . Therefore, the laser beam was not directed vertically so that the geographical coordinates of the aircraft were different from the coordinates of the locations where the laser beam propagated through the Na layer. The

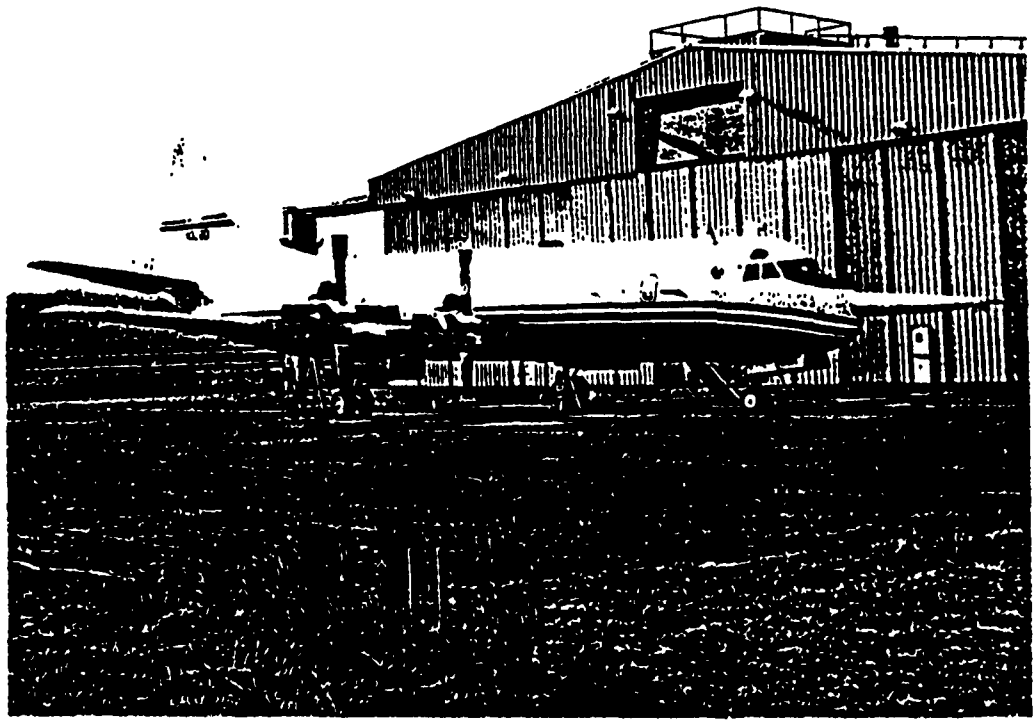


Figure 3.1. A photograph of the NCAR Electra aircraft.

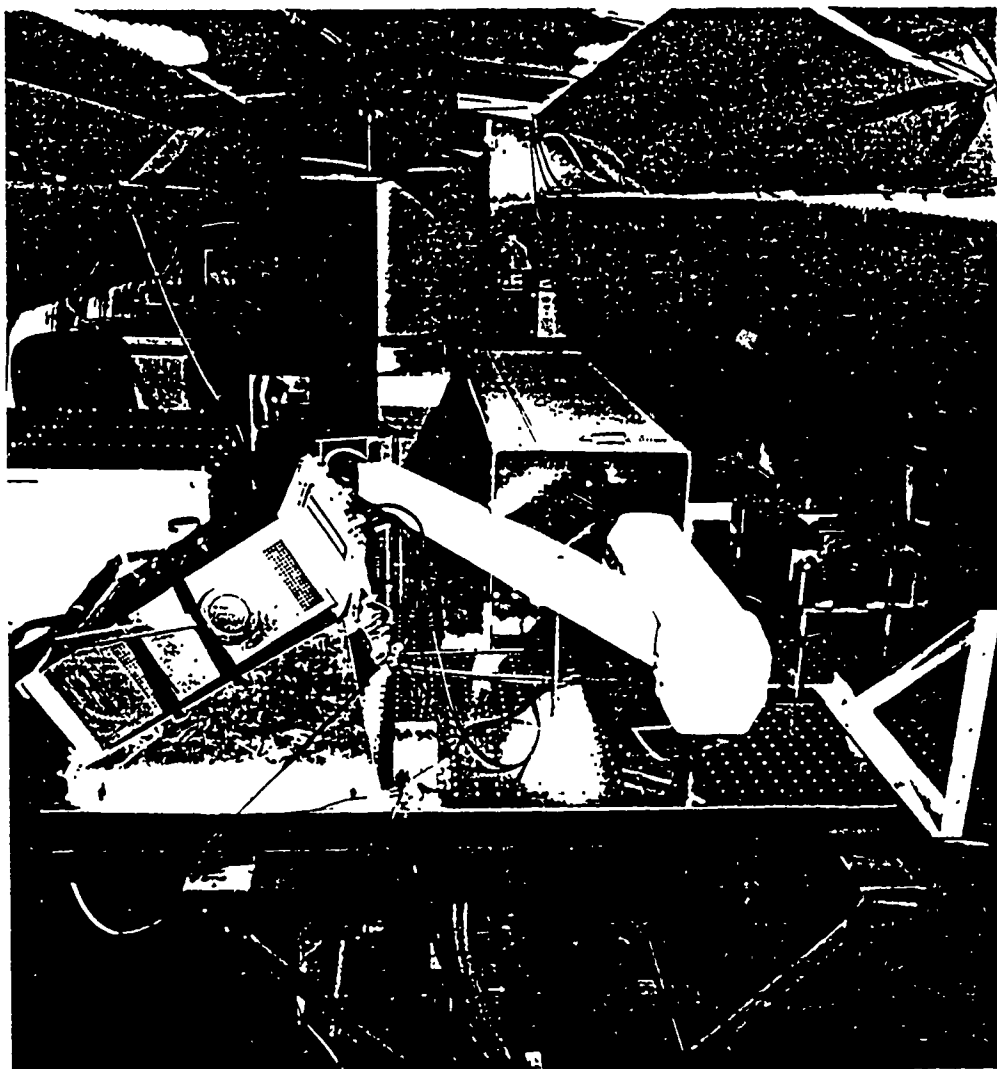


Figure 3.2. A photograph of the interior of the NCAR Electra showing the lidar installation.

geographical coordinates and orientation of the aircraft were used to compute the geographical coordinates of the Na measurements before the Na data were processed.

3.4 Results

In this section, the data collected over the long baselines of the eastward and westward flights will be presented. Both flights included one eastbound and one westbound leg. During the eastward flight, the Na layer appeared to be influenced by quasi-monochromatic waves. To illustrate the effects of the waves, a sequence of Na density profiles is plotted versus longitude in Figure 3.3. These profiles were collected during the eastbound leg. The profiles were filtered vertically with a cutoff of 4 km and horizontally with a cutoff of 70 km. The profiles were then normalized, so that each had the same column abundance. The wavelike features in the layer appear to be repetitive. For example, the profiles near 99-100°W are similar to those near 91-92°W. The horizontal separation distance between these two groups of the profiles is approximately 650 km and appears to be related to the horizontal wavelength of a dominant wave.

The vertical and longitudinal variations of the relative density perturbations also exhibited wavelike features. Figure 3.4 shows the relative density perturbations computed at each altitude from 90 to 97 km at increments of 1 km. Vertical and longitudinal progressions of wavelike features are clearly seen near 99°W and 95°W.

The horizontal wavenumber spectrum corresponding to the profiles from the eastbound leg (Figure 3.3) is plotted in Figure 3.5. The profiles were first filtered vertically with a cutoff of 1 km, and then the horizontal wavenumber spectrum was computed at each altitude from the bottom to the top of the layer at increments corresponding to the vertical resolution of the lidar (i.e., 150 m). The spectrum of Figure 3.5 is the average of the spectra computed from the bottom to the top of the layer. The horizontal wavenumber is the inverse of longitudinal distance along the eastbound leg. In order to estimate the spectral slope, a linear regression fit was

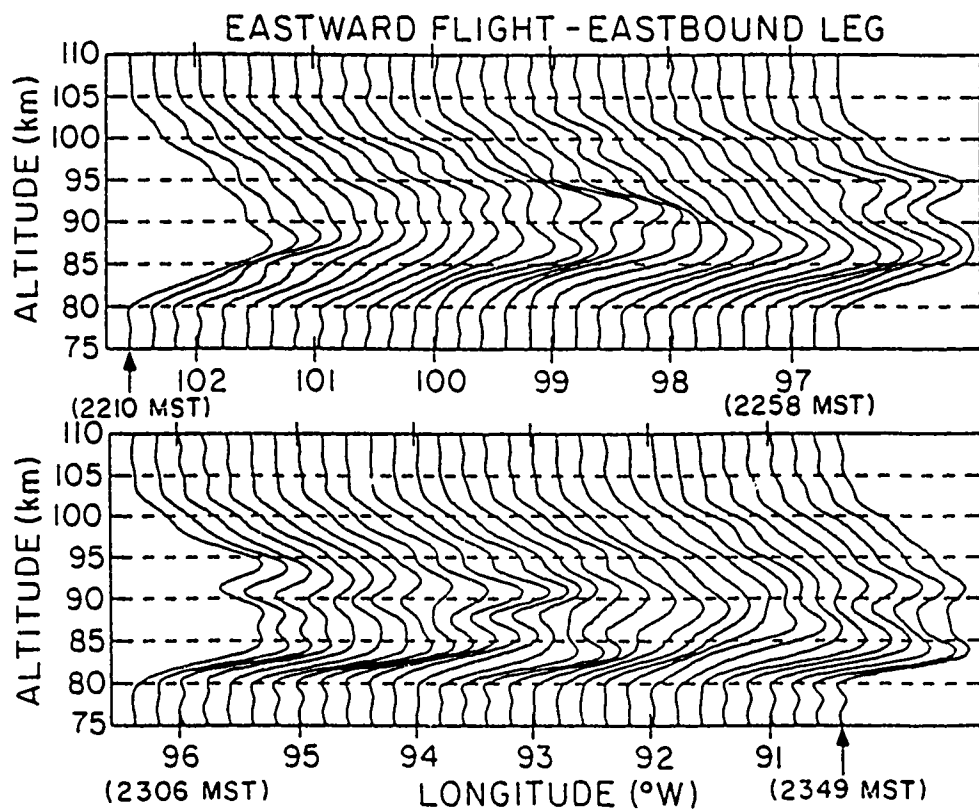


Figure 3.3. Sodium density profiles collected during the eastbound leg of the eastward flight on November 15, 1986. The profiles were filtered vertically with a cutoff of 4 km and horizontally with a cutoff of 70 km. The density profiles were also normalized so that each had the same column abundance and were plotted on a linear scale.

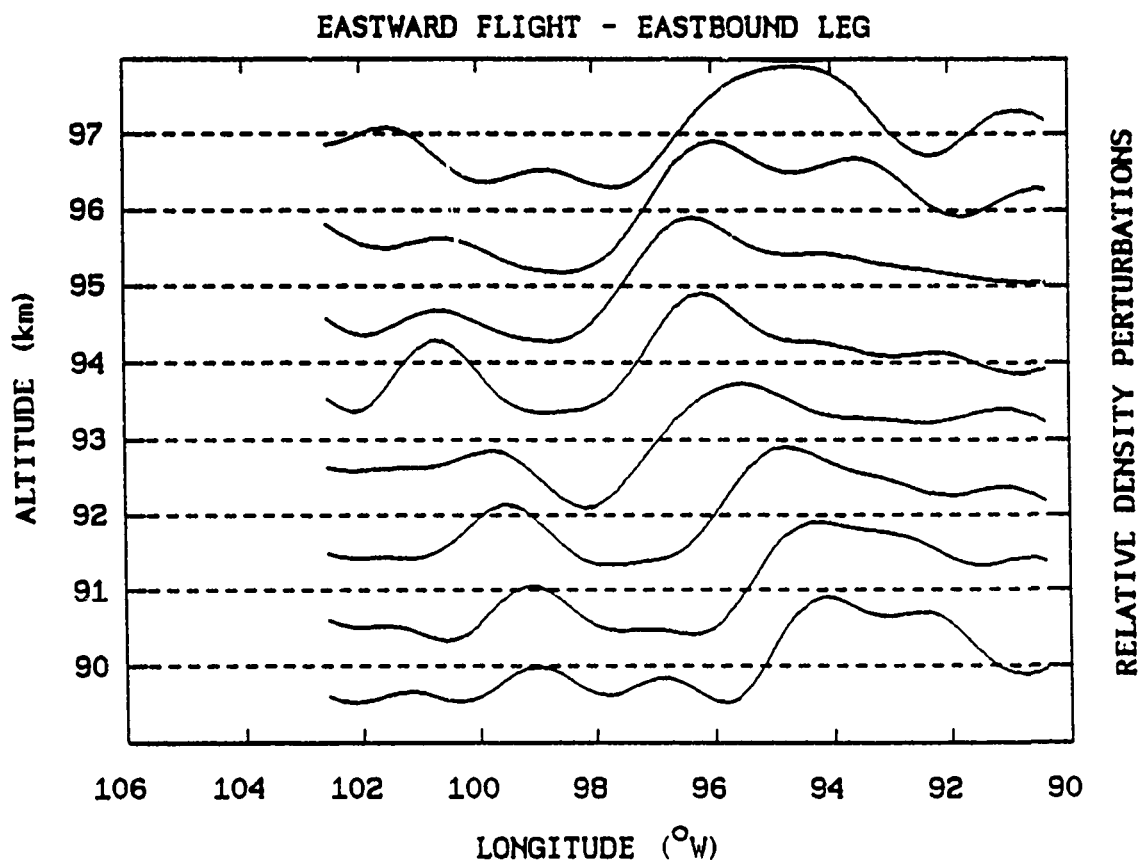


Figure 3.4. Vertical and longitudinal variations of the relative density perturbations computed from the Na lidar data collected during the eastbound leg of the eastward flight on November 15, 1986. Before the data were computed, the density profiles were filtered vertically with a cutoff of 1 km and horizontally with a cutoff of 70 km. Then the relative density perturbations were computed and filtered horizontally with a cutoff of 180 km.

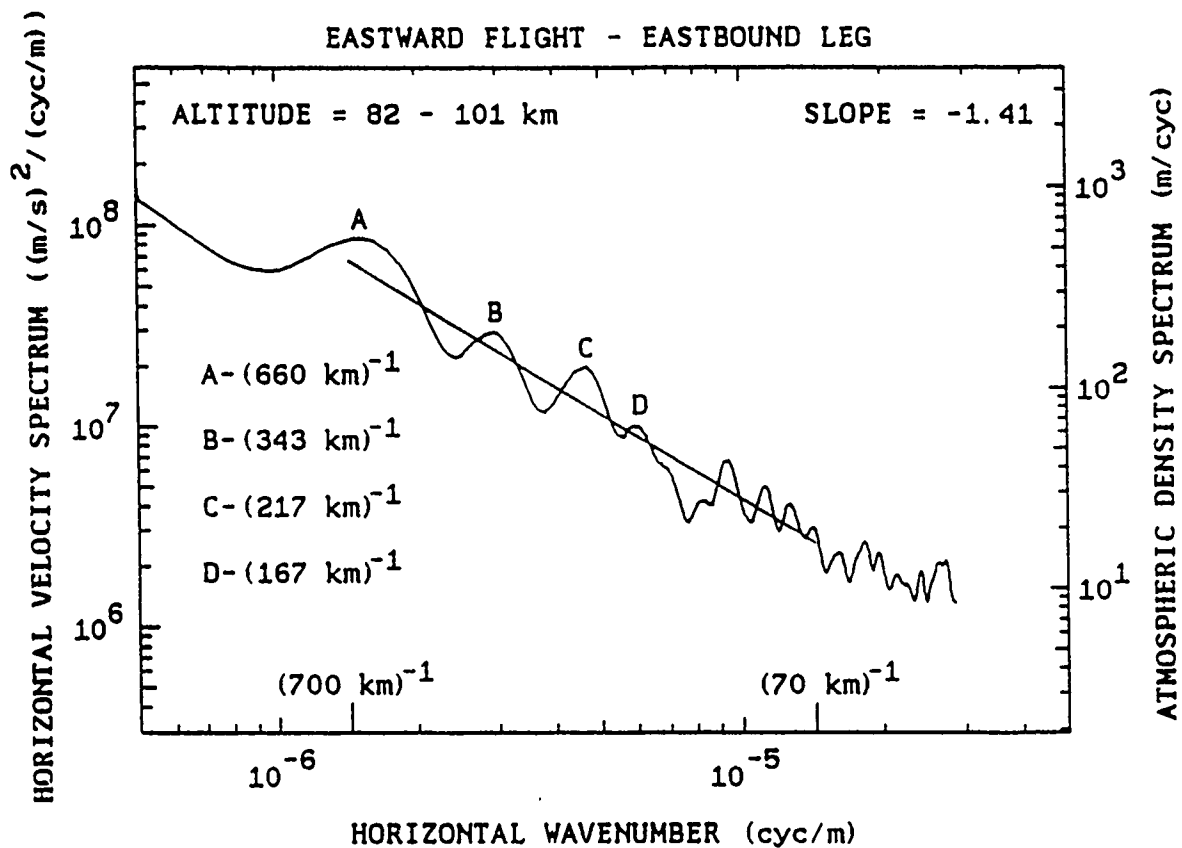


Figure 3.5. Horizontal wavenumber spectrum for the eastbound leg of the eastward flight on November 15, 1986. The data were filtered vertically with a cutoff of 1 km before the spectrum was computed. The straight line is a linear regression fit which was used to estimate the spectral slope over horizontal scales from 70 to 700 km.

performed over the wavenumber range from 1.43×10^{-6} to $1.43 \times 10^{-5} \text{ m}^{-1}$, which corresponds to horizontal scales ranging from 70 to 700 km. The straight line drawn in Figure 3.5 is the linear fit, and the slope is approximately -1.41.

The average vertical wavenumber spectrum corresponding to the profiles of the eastbound leg is plotted in Figure 3.6. The Na density profiles were first filtered horizontally with a cutoff of 70 km, and then the vertical wavenumber spectrum was computed. The shot noise level was estimated and subtracted from the spectrum, and the results were then plotted in Figure 3.6. The straight line is the linear regression fit. The spectral slope is estimated to be -2.80 over the wavenumber range from 10^{-4} to $5 \times 10^{-4} \text{ m}^{-1}$, which corresponds to vertical scales ranging from 2 to 10 km.

A sequence of density profiles collected during the westbound leg of the eastward flight is plotted versus longitude in Figure 3.7. The data were processed in the same manner as those plotted in Figure 3.3. At the flight altitude of approximately 8 km, the prevailing winds were eastward. The ground speed of the aircraft was much slower during the westbound leg. The ground speeds were approximately 181 m s^{-1} eastbound and 121 m s^{-1} westbound. Also, the baseline of the westbound leg was slightly longer, because the westbound flight path extended past Denver and over the front range of the Rockies. Therefore, about twice as many profiles were collected during the westbound leg. In Figure 3.7, the last profile near 90°W was the first observed profile of the westbound leg, because the aircraft was proceeding westward. Note that the features of the profiles observed on the westbound leg near 105°W (Figure 3.7) are quite similar to those observed on the eastbound leg near 99°W (Figure 3.3). These features appear to be propagating westward.

The westward propagating features are also seen in the vertical and longitudinal variations of the relative density perturbations computed from the Na profiles measured on the westbound leg and plotted in Figure 3.8. The maxima of the perturbations near 96°W in Figure 3.4 appear to have propagated westward to 98° in Figure 3.8.

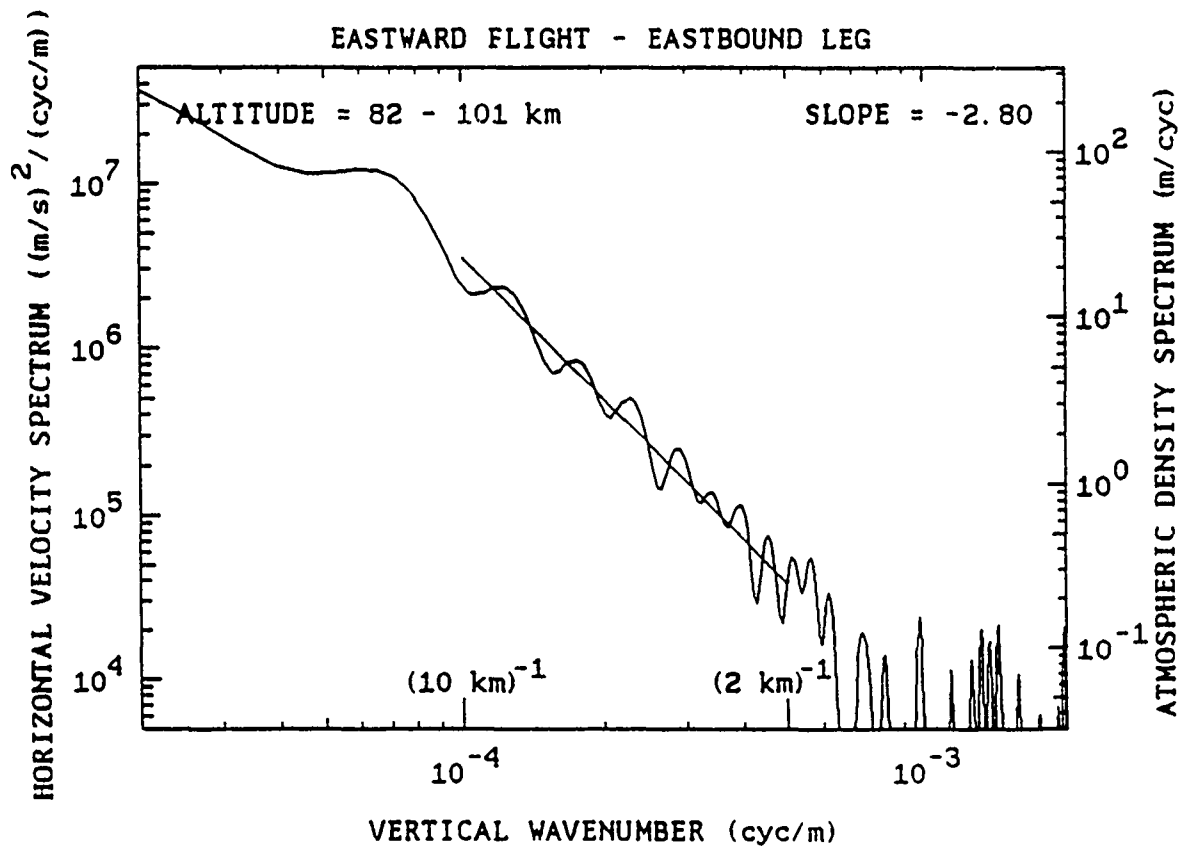


Figure 3.6. Vertical wavenumber spectrum for the eastbound leg of the eastward flight on November 15, 1986. The data were filtered horizontally with a cutoff of 70 km before the spectrum was computed. Then the shot noise level was estimated and subtracted from the spectrum before the spectral slope was estimated. The straight line is a linear regression fit which was used to estimate the spectral slope over vertical scales from 2 to 10 km.

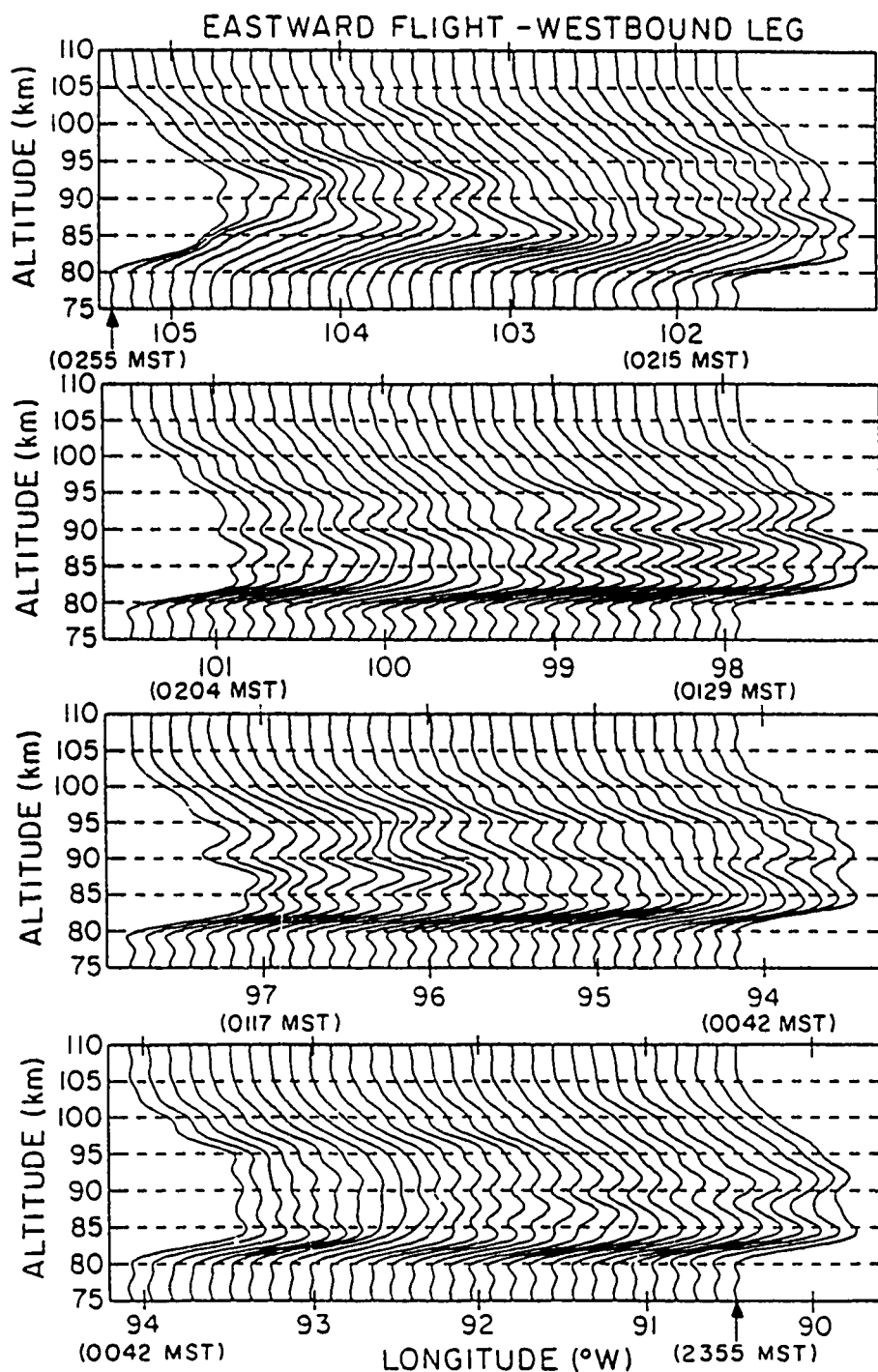


Figure 3.7. Sodium density profiles collected during the westbound leg of the eastward flight on November 15-16, 1986. The profiles were processed in the same manner as those of the eastbound leg of the eastward flight plotted in Figure 3.3.

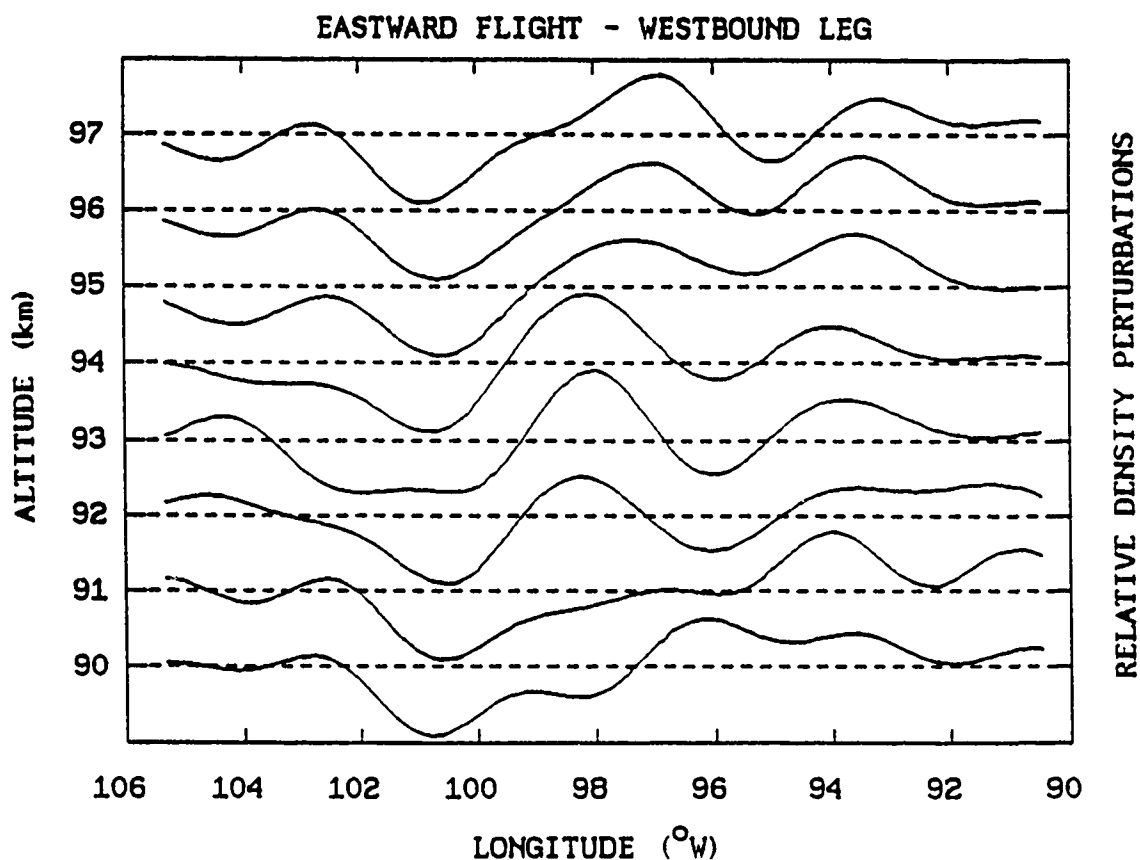


Figure 3.8. Vertical and longitudinal variations of the relative density perturbations computed from the Na lidar data collected during the westbound leg of the eastward flight on November 15-16, 1986. To compute these relative density perturbations, the Na profiles were first filtered vertically with a cutoff of 1 km and horizontally with a cutoff of 70 km. Then the relative density perturbations were computed and filtered horizontally with a cutoff of 233 km.

The horizontal wavenumber spectrum corresponding to the profiles of the westbound leg is plotted in Figure 3.9. The profiles were again vertically filtered with a cutoff of 1 km before the spectrum was computed. Comparison of the spectra of Figure 3.9 and Figure 3.5 reveals two main features. First, the amplitudes of the spectral peaks for wavenumbers lower than $7 \times 10^{-6} \text{ m}^{-1}$ are in general larger for the eastbound observations (Figure 3.5) than for the westbound observations (Figure 3.9). On the other hand, the amplitudes of the peaks for wavenumbers higher than $7 \times 10^{-6} \text{ m}^{-1}$ are smaller for the eastbound observations (Figure 3.5) than for the westbound observations (Figure 3.9). Second, the spectral slope for the westbound observations is estimated to be -1.12 at horizontal scales from 70 to 700 km, which is much smaller than the slope for the eastbound observations. The differences in the spectral amplitudes and slopes could be related to the Doppler effects. For example, the spectral energy corresponding to large scale westward propagating waves would be Doppler shifted to lower wavenumbers for the westbound observations in Figure 3.9, while energy corresponding to small scale eastward propagating waves could be shifted to higher wavenumbers for the westbound observations. The result would be a lower spectral slope for the westbound observations compared to the eastbound observations.

The westward propagating waves are also seen in the wavelike variations of the centroid height of the Na layer plotted in Figure 3.10. Notice that the centroid maxima and minima measured on the westbound leg are separated by larger distances than those measured on the eastbound leg. It appears that the waves which induced the centroid height perturbations were propagating westward. In fact, the centroid maxima and minima measured on the eastbound leg can be matched with those measured on the westbound leg. The corresponding maxima and minima are marked in Figure 3.10. From the measurement times and horizontal separation distances between the corresponding maxima (or minima), westward velocities were computed and summarized in Table 3.1.

It appears that at least two different waves were dominating the Na layer during the

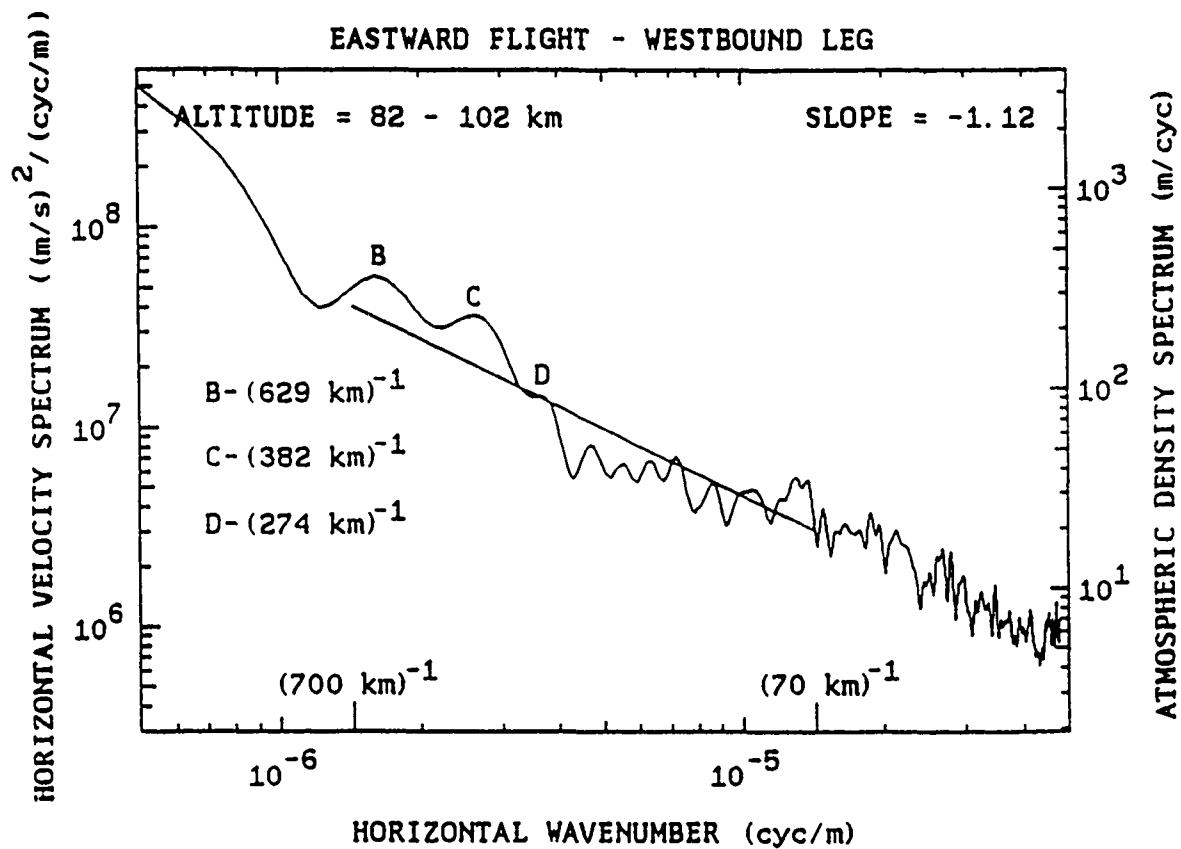


Figure 3.9. Horizontal wavenumber spectrum for the westbound leg of the eastward flight on November 15-16, 1986. The data were processed in the same manner as the data of the eastbound leg of the eastward flight.

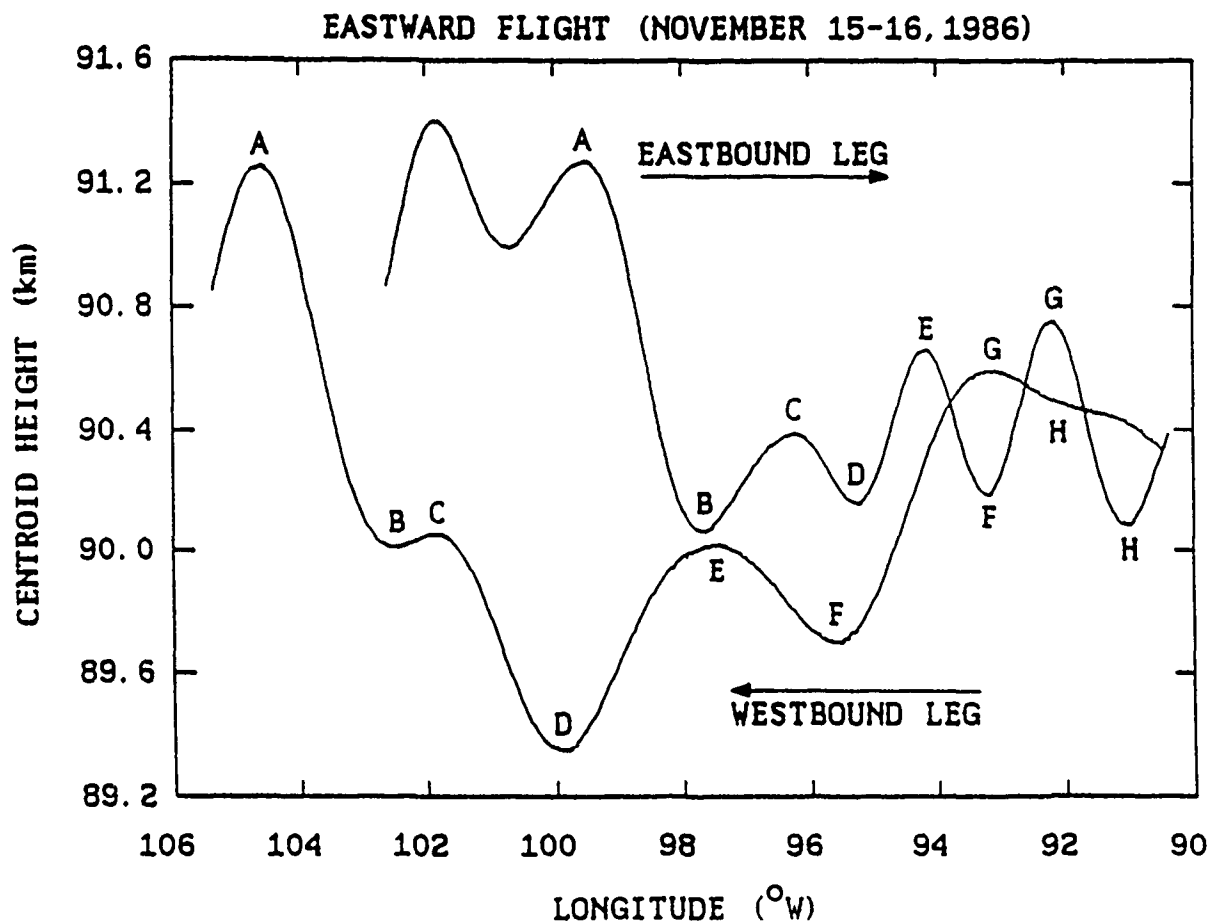


Figure 3.10. The longitudinal variations of the centroid height of the Na layer observed during the eastward flight on November 15-16, 1986.

Table 3.1. Horizontal Velocities of the Maxima and Minima of the Centroid Height of the Na Layer Observed During the Eastward Flight on November 15-16, 1986

		Eastbound Leg		Westbound Leg		Westward Velocity (m s^{-1})
	Feature	Longitude ($^{\circ}\text{W}$)	Time (MST)	Longitude ($^{\circ}\text{W}$)	Time (MST)	
A	Maximum	99.53	22:38	104.60	02:46	29.6
B	Minimum	97.67	22:53	102.52	02:22	33.6
C	Maximum	96.27	23:04	101.88	02:14	42.9
D	Minimum	95.27	23:12	99.92	01:51	42.6
E	Maximum	94.20	23:20	97.52	01:22	39.3
F	Minimum	93.20	23:27	95.64	01:01	37.6
G	Maximum	92.20	23:35	93.16	00:32	24.5

flight. One wave appeared to have a much shorter zonal wavelength than the other. The zonal wavelength of the shorter wave is related to the separation distance between two centroid maxima, C and E, which was approximately 180 km on the eastbound leg and 380 km on the westbound leg. The wavelength of the longer wave is related to the distance between centroid maxima, A and G, which was 640 km on the eastbound leg and about 1000 km on the westbound leg. The horizontal wavenumber spectra for the eastbound and westbound observations exhibited spectral peaks corresponding to these two waves. The shorter wavelength wave corresponds to spectral peaks C in Figures 3.5 and 3.9, and the longer wavelength wave corresponds to spectral peak A in Figure 3.5. The spectral peak for the longer wavelength wave could not be seen in the spectrum for the westbound observations, because the Doppler shifted zonal wavelength of this wave was too close to the flight leg distance of 1295 km. However, the spectral amplitudes at the wavenumbers near $(1000 \text{ km})^{-1}$ appear to be much larger in Figure 3.9 for the westbound leg compared to the eastbound leg.

There are four distinct spectral peaks, A through D, in the horizontal wavenumber spectrum for the eastbound observations (Figure 3.5), and three distinct peaks, B through D, in the horizontal wavenumber spectrum for the westbound observations (Figure 3.9). Spectral peaks B and D in Figure 3.5 appear to be sidelobes of spectral peaks A and C, respectively. Similarly, spectral peak D in Figure 3.9 appears to be a sidelobe of spectral peak C, and spectral peak B also appears to be a sidelobe which corresponds to the longer wavelength wave. The wavenumber separation between the main spectral peak and the first sidelobe of a monochromatic wave is related to the horizontal extent of the observations by

$$\Delta k = \frac{1.5}{\Delta x} \quad (3.40)$$

where Δk is the wavenumber separation between the main peak and the first sidelobe.

The horizontal extent, Δx , for the eastbound observations was 1060 km, and that for the westbound observations was 1300 km. By using Equation (3.40), the wavenumber separation

between the main peak and the first sidelobe is calculated to be $1.4 \times 10^{-6} \text{ m}^{-1}$ for the eastbound observations, and $1.2 \times 10^{-6} \text{ m}^{-1}$ for the westbound observations. The wavenumber separation between spectral peaks A and B in the spectrum for the eastbound observations (Figure 3.5) is $1.4 \times 10^{-6} \text{ m}^{-1}$, and the separation between spectral peaks C and D is again $1.4 \times 10^{-6} \text{ m}^{-1}$. The wavenumber separation between spectral peaks C and D in the spectrum for the westbound observations (Figure 3.9) is 10^{-6} m^{-1} .

The observed Doppler shifted wavelengths of the two waves are related to the intrinsic zonal wavelengths by the following equation.

$$\lambda_{in} = \lambda_{ob} \left(1 - \frac{V_{ob}}{V_a} \right) \quad (3.41)$$

where λ_{in} = intrinsic wavelength along the flight path,

λ_{ob} = observed wavelength along the flight path,

$V_{ob} = V_p + V_b$ = observed phase velocity along the flight path,

V_p = intrinsic phase velocity along the flight path,

V_b = background atmospheric wind velocity along the flight path at the altitudes of the Na layer, and

V_a = aircraft velocity.

The background wind velocity was measured at Platteville, Colorado with an ST radar during the period from November 4 to 20, 1986. The average zonal wind velocity was about 5 m s^{-1} eastward over the altitude range of the Na layer (see Section 4.3).

Because observations were made over both the eastbound and westbound flight legs, and the aircraft velocity is known, Equation (3.41) can be solved for both the intrinsic wavelength and observed phase velocity along the flight path. The intrinsic wavelength along the flight path is

$$\lambda_{in} = \frac{\lambda_{ob}^e \lambda_{ob}^w (V_a^w - V_a^e)}{\lambda_{ob}^e V_a^w - \lambda_{ob}^w V_a^e} \quad (3.42)$$

where λ_{ob}^e = wavelength observed during the eastbound flight leg,
 λ_{ob}^w = wavelength observed during the westbound flight leg,
 V_a^e = average aircraft velocity during the eastbound flight leg, and
 V_a^w = average aircraft velocity during the westbound flight leg.

Similarly, the observed phase velocity along the flight path is

$$V_{ob} = \frac{V_a^e V_a^w (\lambda_{ob}^e - \lambda_{ob}^w)}{\lambda_{ob}^e V_a^w - \lambda_{ob}^w V_a^e} \quad (3.43)$$

By using the two wavelengths of spectral peaks C from Figure 3.5 and 3.9, the intrinsic zonal wavelength of the shorter wavelength wave is computed to be 263 km. The observed phase velocity is approximately 38 m s⁻¹ westward, which is quite comparable to the westward velocities of centroid maxima, C and E, and minima, D and F listed in Table 3.1. By subtracting the background wind velocity from the observed phase velocity, the intrinsic zonal phase velocity is calculated to be 43 m s⁻¹ westward. The intrinsic wave period is calculated to be 102 min.

From ground-based Na lidar observations obtained at Urbana, Illinois, *Gardner and Voelz* [1987] found surprisingly systematic relationships between horizontal and vertical wavelengths and the observed periods of quasi-monochromatic waves. According to the *Gardner and Voelz* [1987] empirical relationships, an observed period of 102 min corresponds to a horizontal wavelength of 101 km and a vertical wavelength of 4.6 km. The zonal wavelength will be larger than the intrinsic horizontal wavelength if the wave is not propagating zonally. The relationship between the intrinsic horizontal wavelength and zonal wavelength is

$$\lambda_{zonal} = \frac{\lambda_x}{\cos \alpha} \quad (3.44)$$

where α is the angle between the wave propagation direction and the east-west line. By assuming the intrinsic horizontal wavelength is approximately equal to the value predicted by the *Gardner and Voelz* [1987] empirical relationship, the angle, α , is calculated from Equation

(3.44) to be approximately 70° . Thus the wave appears to be propagating at an approximate azimuth angle of either 200° or 340° .

The parameters of the longer wavelength wave can be also estimated by using Equations (3.42) and (3.43). The observed wavelength for the eastbound observations was 660 km, which is the wavelength of spectral peak A in the horizontal wavenumber spectrum (Figure 3.5). For the westbound observations, the wavelength is estimated from the centroid height data to be approximately 1030 km. By using these two horizontal wavelengths, the intrinsic zonal wavelength is calculated to be 772 km, and the observed phase velocity is approximately 30 m s^{-1} westward. This velocity is comparable to the westward velocities of centroid maxima A and G and minimum B (see Table 3.1). The intrinsic period is computed to be 363 min or 6.1 hours. This period of 6.1 hours strongly suggests that the wave is a westward propagating tide. The vertical wavelength of this wave is related to the intrinsic horizontal wavelength by the gravity wave dispersion relation,

$$\lambda_z = \frac{T_B}{T_{in}} \lambda_x \quad (3.45)$$

where T_B = Brunt-Vaisala period (5 min),

T_{in} = intrinsic wave period.

The intrinsic horizontal wavelength of the wave was not measured during the flight. However, by using the intrinsic zonal wavelength of 772 km and the intrinsic period of 6.1 hours, the vertical wavelength is calculated to be 10.6 km. From the systematic relationships obtained from the ground-based Na lidar observations by *Gardner and Voelz* [1987], an observed period of 6.1 hours would correspond to a horizontal wavelength of 734 km and a vertical wavelength of 9.8 km. These horizontal and vertical wavelengths are quite comparable to those calculated from the airborne data, and provide further support for the assumption that this wave was propagating approximately due west. The theoretical results of *Gardner and Shelton* [1985] and *Gardner et al.* [1986] indicate that the Na layer centroid height perturbations are largest for waves with

vertical wavelengths between 10 and 20 km and for vertical wavelengths greater than 35 km. The peak-to-peak centroid height variations caused by the 6.1-hour period wave are almost 1 km for the westbound flight leg. The Platteville, Colorado radar also measured a strong 6-hour period wind oscillation near 90 km during the period November 4 to 20, 1986 (see Section 4.3).

The wave parameters including horizontal wind amplitudes corresponding to spectral peaks A and C are summarized in Table 3.2. The horizontal wavelengths of the waves observed with the ground-based Na lidar [*Gardner and Voelz, 1987*] are plotted versus period in Figure 3.11. The intrinsic zonal wavelengths and intrinsic periods measured during the eastward flight are also plotted in Figure 3.11. The straight line is a maximum likelihood linear regression fit which was used to estimate the slope for the ground-based observations. As discussed earlier, the intrinsic zonal wavelength of the 6.1-hour period wave appears to be quite comparable to the horizontal wavelengths estimated from the ground-based observations for the same period. The wave corresponding to spectral peak C appears to have propagated at an azimuth angle of approximately 200° or 340° . The horizontal distance between the location of the airborne lidar observations and the location of the source of a wave can be estimated roughly by multiplying the height of the Na layer by the ratio of the intrinsic horizontal wavelength to the vertical wavelength of the observed wave. For the wave corresponding to spectral peak C, the average horizontal distance to the tropospheric source was computed to be about 2000 km. Interestingly, during the period from November 13 to 19, a strong meteorological low pressure system developed over northern Hudson Bay, Canada. The azimuth angle of a line drawn from northern Hudson Bay to the average horizontal location of the flight track is about 200° , and the horizontal distance is about 2500 km. Although it is speculative, this large stationary low pressure system might be responsible for the wave corresponding to spectral peak C.

The kinetic energy distribution for the waves observed with the airborne Na lidar is plotted versus temporal frequency in Figure 3.12 along with the ground-based data reported by *Gardner and Voelz [1987]*. The kinetic energy for the airborne measurements was calculated by

**Table 3.2. Monochromatic Wave Parameters Estimated
from the Horizontal Wavenumber Spectra for the
Eastward Flight on November 15-16, 1986**

Spectral Peak	A	C
Observed Zonal Wavelength		
Eastbound Observations (km)	660	217
Westbound Observations (km)	1030 ^a	382
Intrinsic Zonal Wavelength (km)	772	263
Observed Zonal Phase Velocity (m s ⁻¹ westward)	30	38
Intrinsic Zonal Phase Velocity (m s ⁻¹ westward) ^b	35	43
Intrinsic Period (min)	363	102
Horizontal Wind Amplitude		
Eastbound Observations (m s ⁻¹)	12.7	6.1
Westbound Observations (m s ⁻¹)	NA	7.5

^aEstimated from the centroid height variations measured on the westbound leg.

^bCalculated assuming background wind of 5 m s⁻¹ eastward. This value was the measured average zonal wind at Platteville, Colorado during the period from November 4 - 20, 1986.

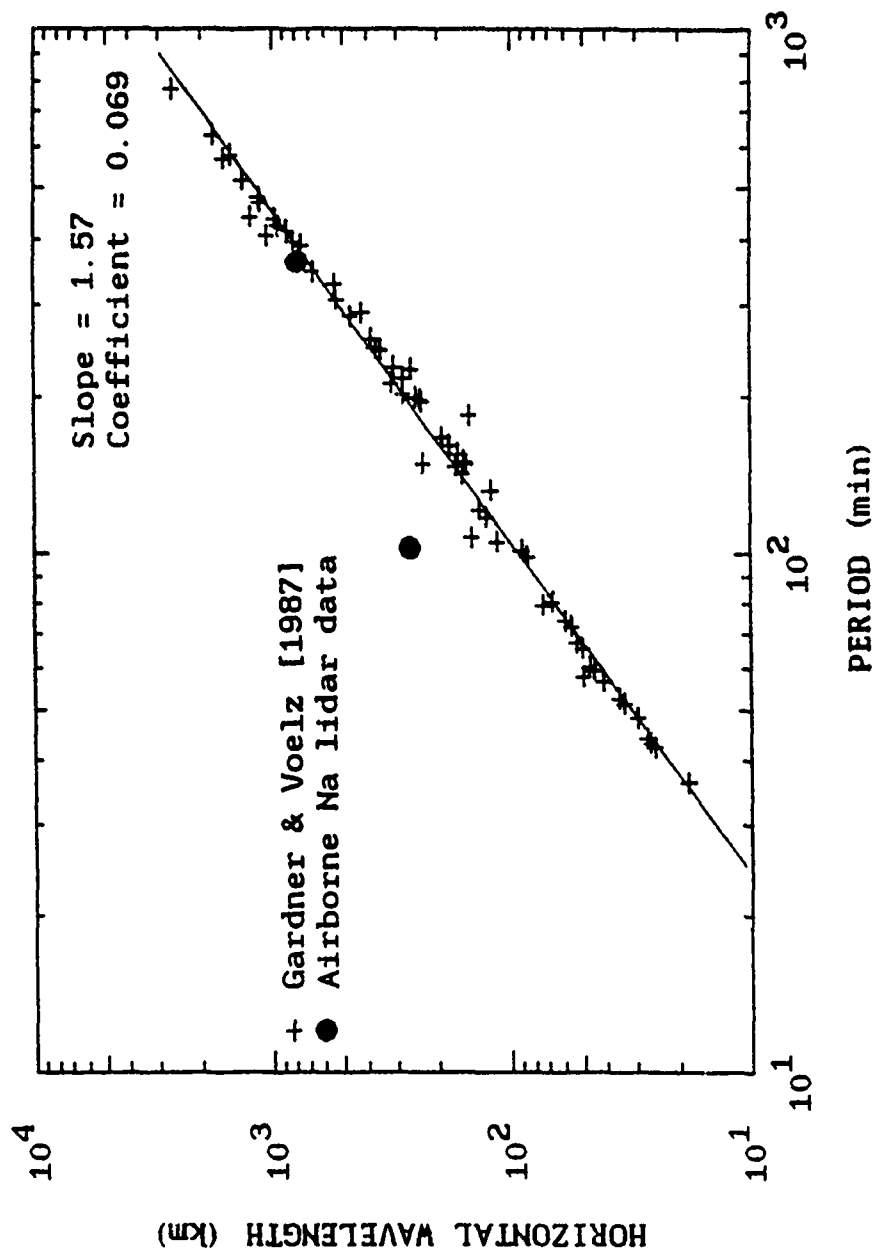


Figure 3.11. The horizontal wavelengths of the waves observed with the ground-based (after Gardner and Voelz, 1987) and airborne Na lidars are plotted versus period. Crosses represent the ground-based observations obtained at Urbana, Illinois during the winter seasons from 1980 to 1986. Circles represent the intrinsic zonal wavelengths and intrinsic periods measured during the eastward flight on November 15-16, 1986. The straight line is a maximum likelihood linear regression fit which was used to estimate the slope for the ground-based observations.

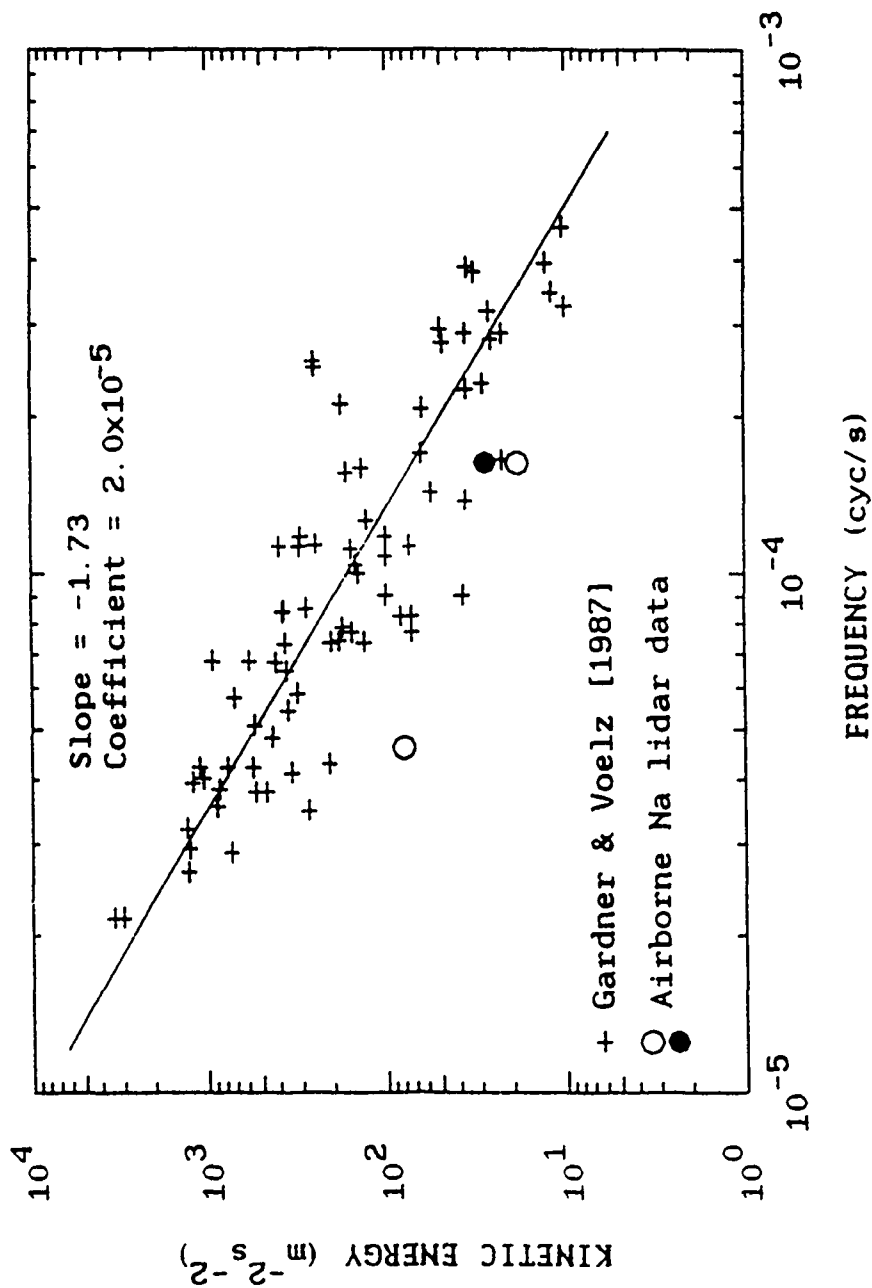


Figure 3.12. The kinetic energy of quasi-monochromatic gravity waves plotted versus temporal frequency. Crosses represent the ground-based Na lidar observations obtained at Urbana, Illinois during the winter seasons from 1980 to 1986 (after Gardner and Voelz, 1987). Circles represent the kinetic energy distribution for the quasi-monochromatic waves measured during the eastward flight on November 15-16, 1986. The shaded circle is the data measured during the westbound leg, and the open circles are the data measured during the eastbound leg. The straight line represents a maximum likelihood linear regression fit for the ground-based measurements.

using Equation (3.39) and assuming that $\beta \Delta z < 1$ for each wave. The kinetic energies of the waves observed with the airborne lidar are slightly lower than those observed with the ground-based lidar.

The average vertical wavenumber spectrum for the westbound leg is plotted in Figure 3.13. The slope is approximately -2.68 at vertical scales from 2 to 10 km, and is shallower than the slope of the vertical wavenumber spectrum measured on the eastbound leg. The vertical wavenumber spectrum for the westbound leg was larger in amplitude than that for the eastbound leg for most wavenumbers. The vertical wavelength of the dominant wave in the spectrum is about 9.6 km, which appears to be the vertical wavelength of the wave with the period of 6.1 hours.

Figures 3.14 and 3.15 show, respectively, the horizontal and vertical wavenumber spectra computed from the data collected during the eastbound and westbound legs of the westward flight. The difference in the horizontal wavenumber spectra plotted in Figures 3.14 a and b may be caused by Doppler effects and localized gravity wave packets. The vertical wavenumber spectra for both flight legs are plotted in Figures 3.15 a and b. The vertical wavenumber spectrum for the westbound leg was smaller in amplitude for most wavenumbers, while the slope for the westbound leg was steeper. The vertical wavelengths of the dominant waves in both vertical wavenumber spectra were again 9.6 km.

In Figure 3.16, the rms horizontal wind velocities measured on both eastward and westward flights are plotted. During the eastward flight, the wind amplitudes of the velocity perturbations were smaller on the eastbound leg compared to amplitudes measured on the westbound leg. However, during the westward flight the wind amplitudes were larger on the eastbound leg. These changes of the amplitudes could be caused by diurnal variations in wind activity and the differences in the measurement times on the two flights. At Urbana, Illinois, the ground-based lidar observations have revealed that the rms horizontal wind velocity typically increases during the night toward the early morning. Therefore, this temporal change could be

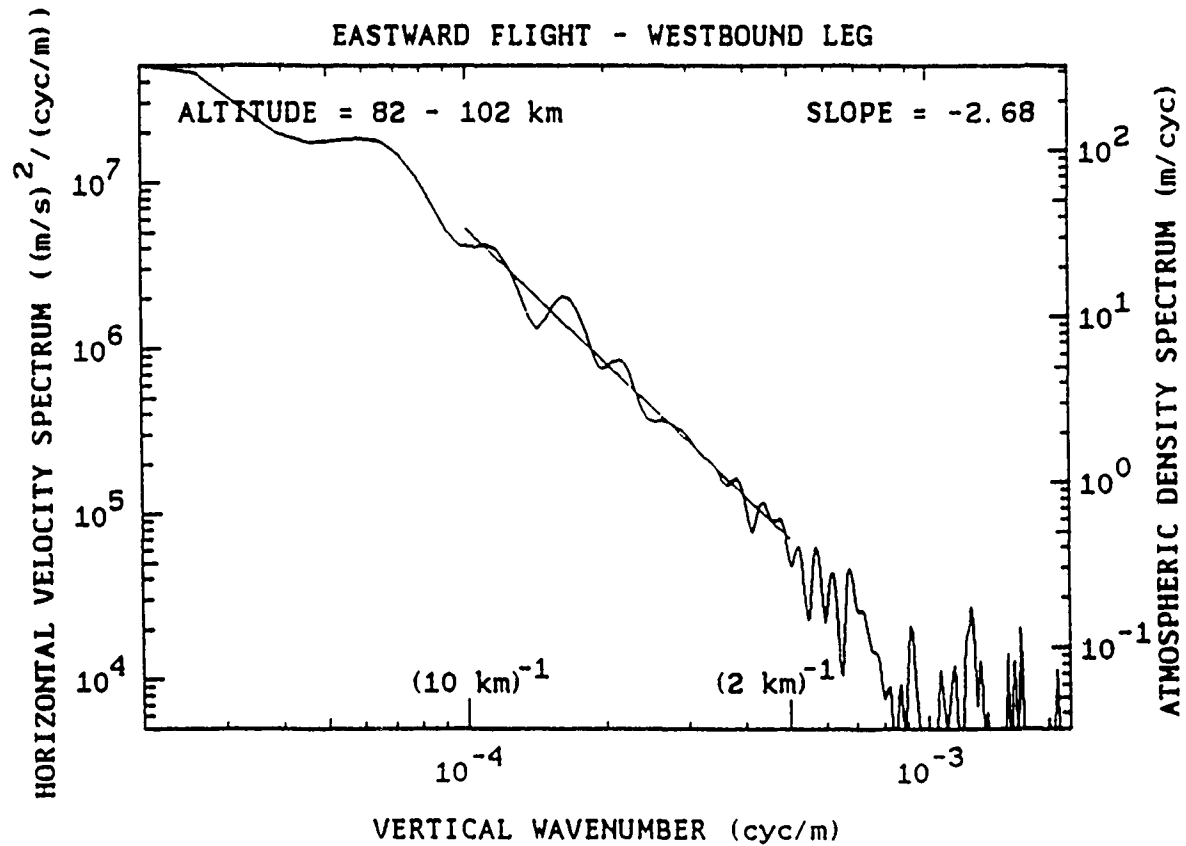


Figure 3.13. Vertical wavenumber spectrum for the westbound leg of the eastward flight on November 15-16, 1986.

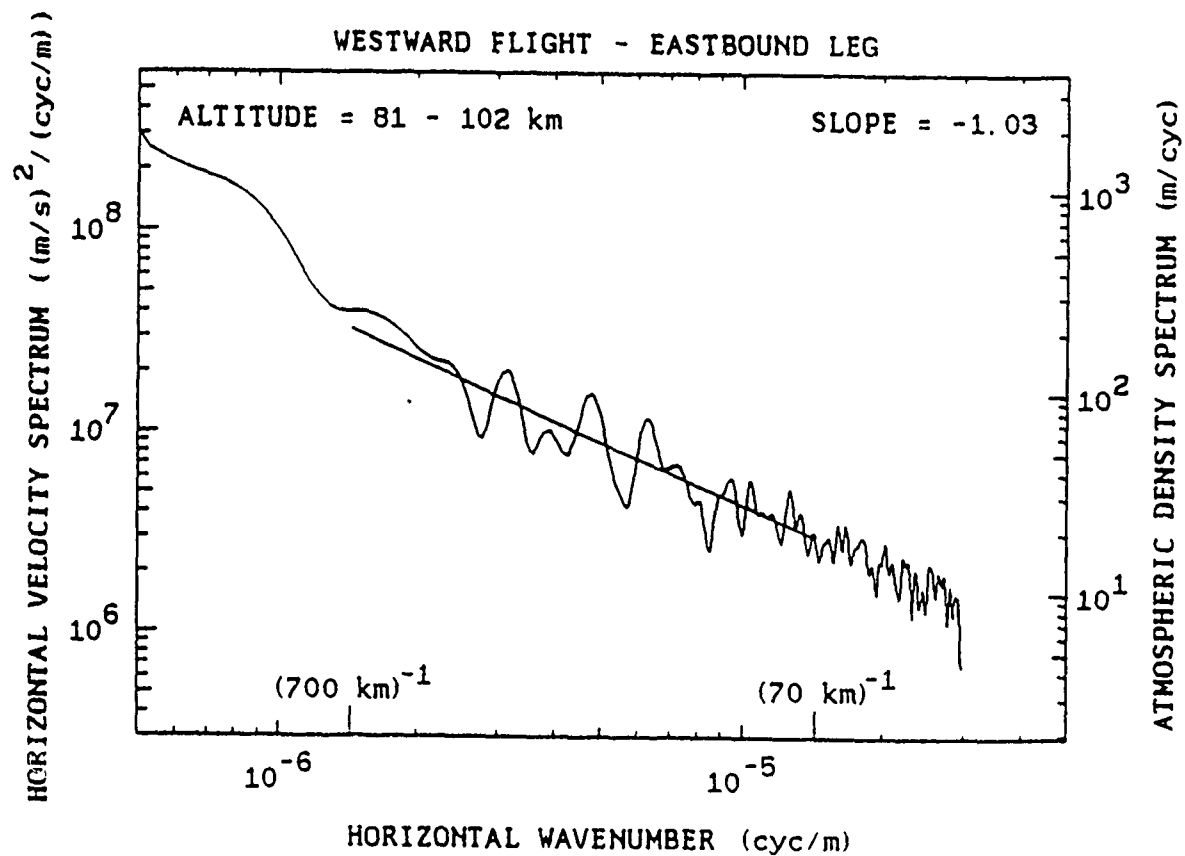


Figure 3.14 a). Horizontal wavenumber spectrum for the eastbound leg of the westward flight on November 17-18, 1986.

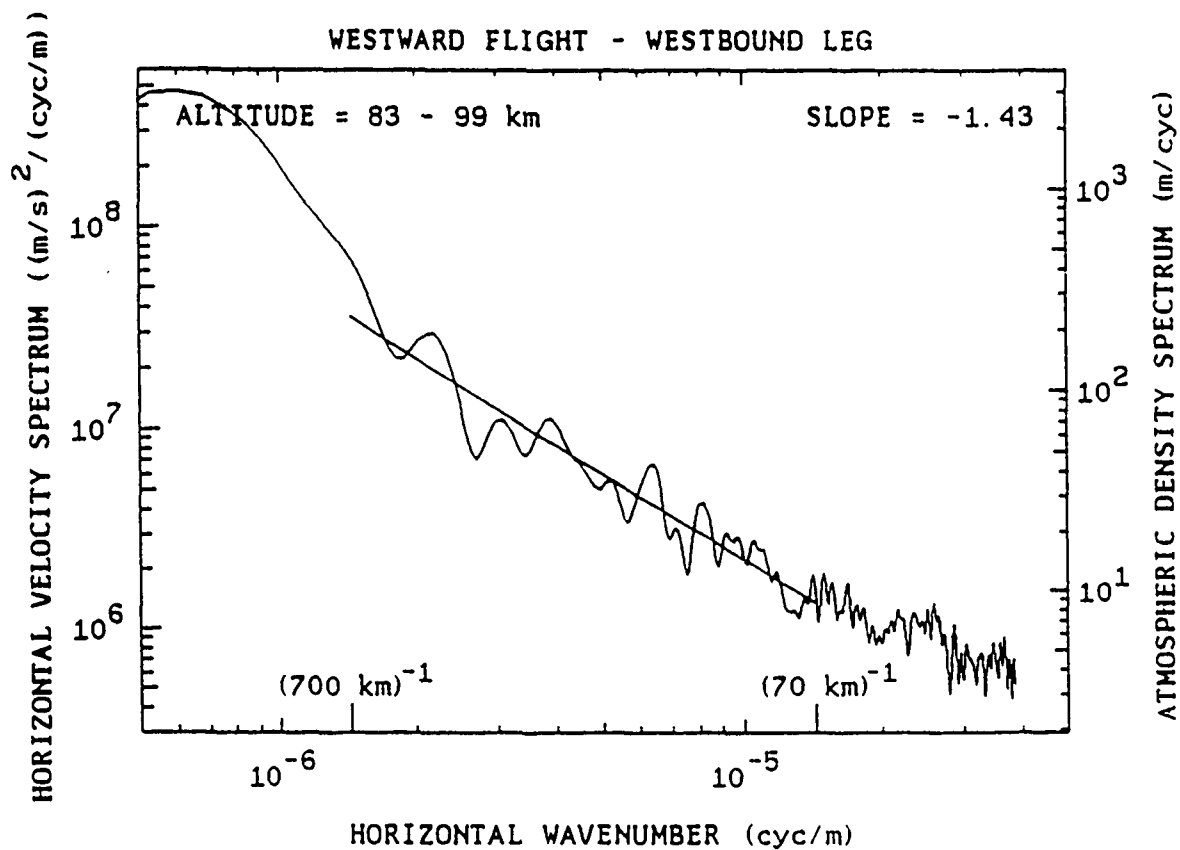


Figure 3.14 b). Horizontal wavenumber spectrum for the westbound leg of the westward flight on November 17-18, 1986.

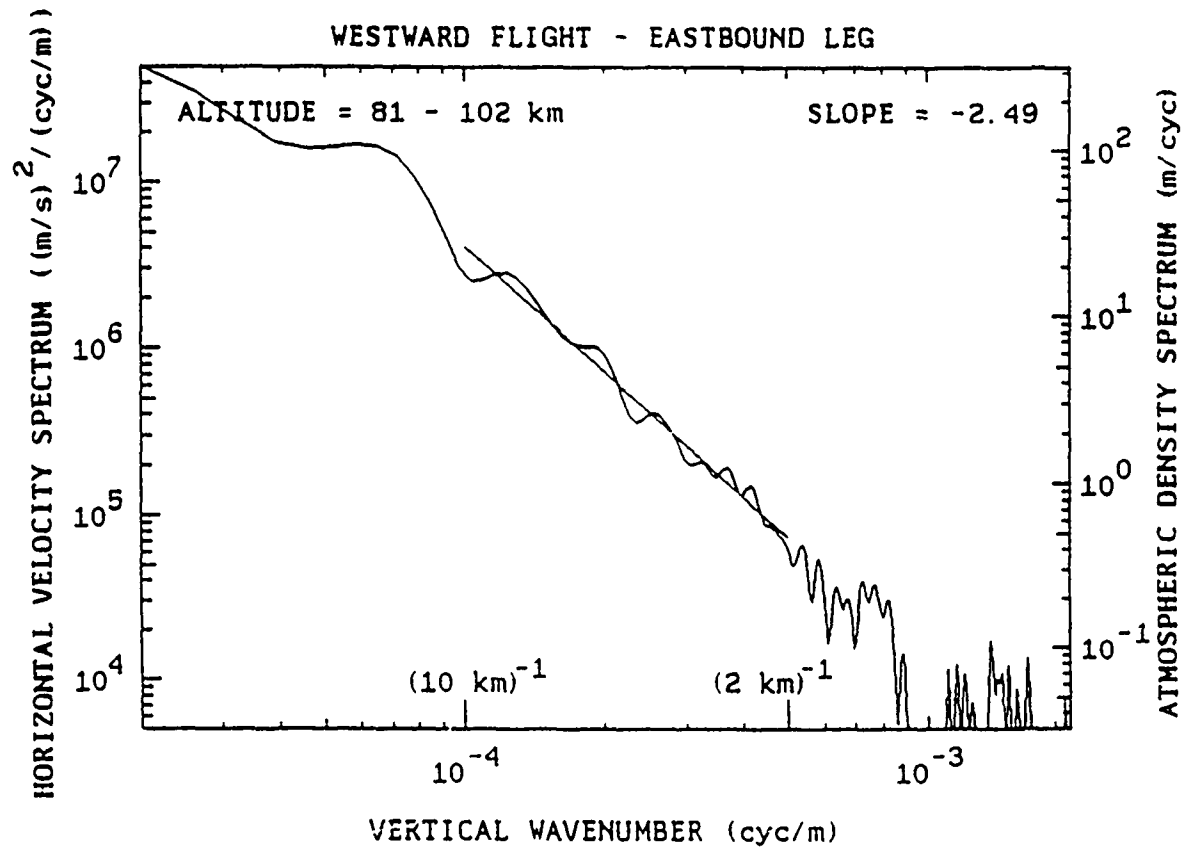


Figure 3.15 a). Vertical wavenumber spectrum for the eastbound leg of the westward flight on November 17-18, 1986.

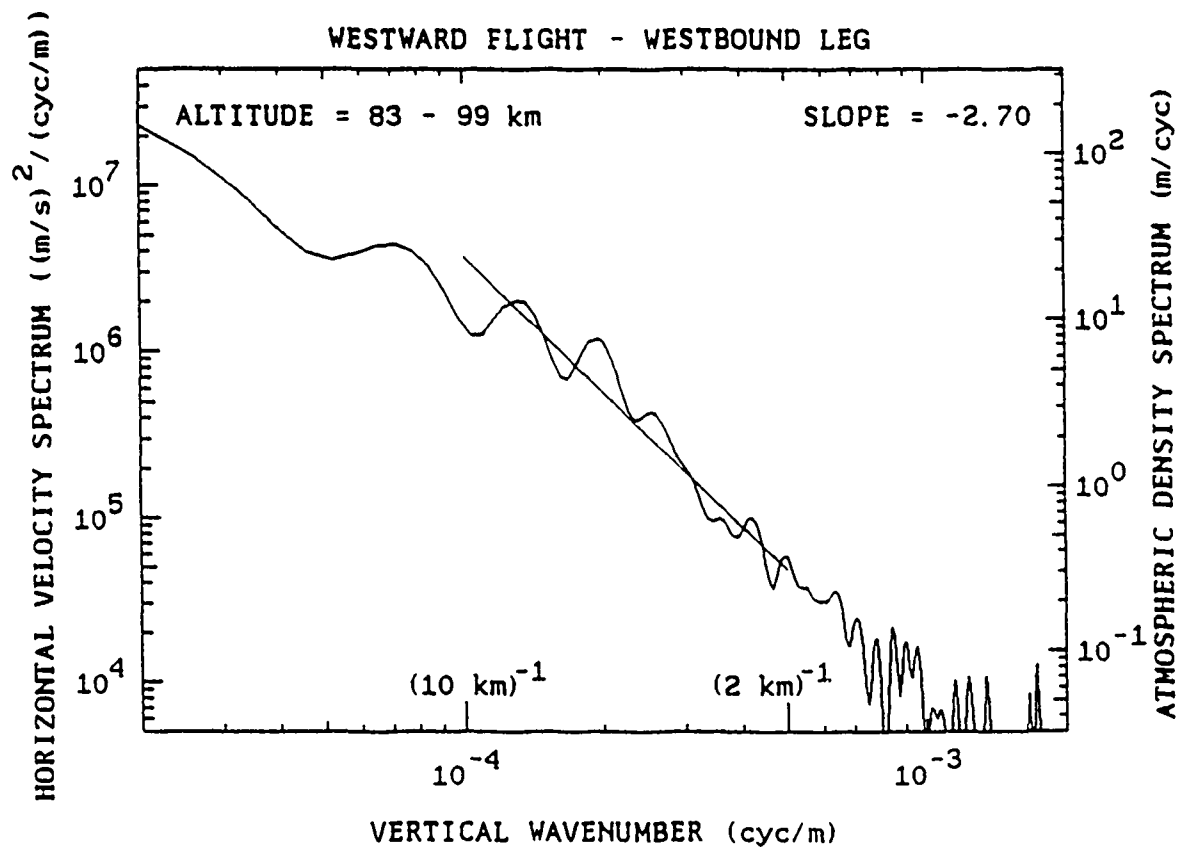


Figure 3.15 b). Vertical wavenumber spectrum for the westbound leg of the westward flight on November 17-18, 1986.

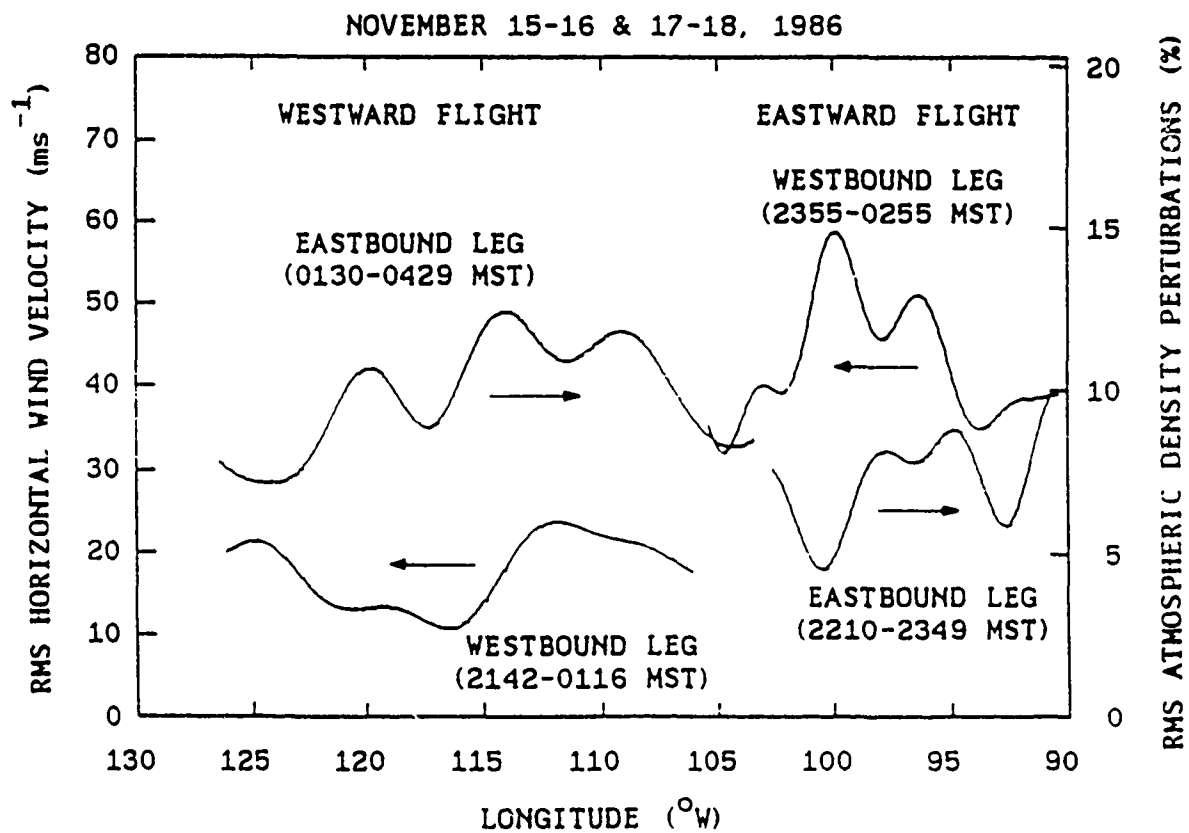


Figure 3.16. The rms horizontal wind velocities inferred from the airborne Na lidar data collected during the eastward (November 15-16, 1986) and westward (November 17-18, 1986) flights.

responsible for the changes of the rms wind velocities observed during the flights. The eastbound leg of the eastward flight and the westbound leg of the westward flight were both conducted during the late evening. Another feature of Figure 3.16 is that the rms wind velocity increased in general with longitude from the Pacific Coast to the Great Plains.

In Figure 3.17, the rms horizontal wind velocities measured with the UIUC ground-based lidar at four different locations along with the velocities measured during the airborne observations are plotted over a longitude range of 70°W to 160°W . The four locations include Mauna Kea Observatory (MKO), Hawaii ($20^{\circ}\text{N}, 155^{\circ}\text{W}$), NCAR in Broomfield, Colorado ($40^{\circ}\text{N}, 105^{\circ}\text{W}$), UIUC, Illinois ($40^{\circ}\text{N}, 88^{\circ}\text{W}$), and Goddard Space Flight Center (GSFC), Maryland ($39^{\circ}\text{N}, 78^{\circ}\text{W}$). The measurements at Mauna Kea Observatory were obtained on three different nights, January 19, 20, and 21, 1987. The three dots near the longitude of 155°W in Figure 3.17 indicate the average values of the rms horizontal wind velocities measured on the three nights, and the lines indicate the ranges of the velocities measured on each night. The measurements at Broomfield were obtained on the nights of November 12 and 19, 1986. The dots near the longitude of 105°W again indicate the average values measured on the two nights, and the lines indicate the ranges of the measured values. The measurements at the Goddard Space Flight Center were obtained on the nights of October 16, 19, and 21, 1981. The dots and lines near the longitude of 78°W indicate the average values and the ranges of the measured values, respectively. The dot near the longitude of 88°W represents the measurements at UIUC where the ground-based lidar observations have been obtained for the last 10 years. The rms horizontal winds measured at UIUC exhibit a strong semi-annual oscillation along with a weak annual oscillation [Gardner and Senft, 1989]. A regression fit was performed to estimate the amplitudes and phases of both the semi-annual and annual oscillations for the data obtained from 1984 to 1986. The dot in Figure 3.17 near 88°W represents the value for the fit corresponding to the middle of November. The error bar indicates the rms difference between the fitted value and the measured values. The magnitudes of the rms horizontal wind velocities measured in Hawaii

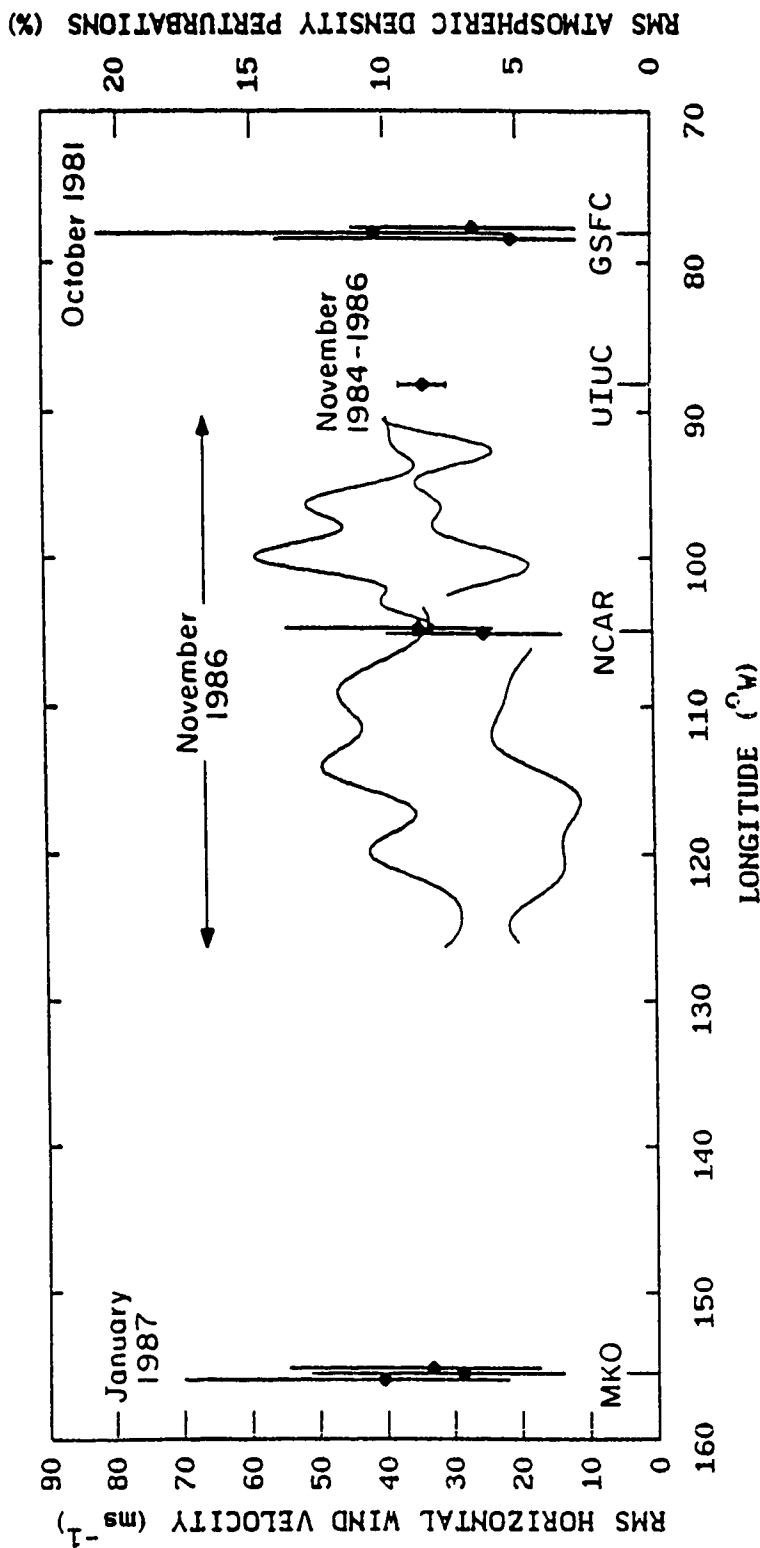


Figure 3.17. The rms horizontal wind velocities inferred from the airborne Na lidar data, and from the ground-based Na lidar data obtained at Mauna Kea Observatory (MKO), Hawaii (20°N, 155°W), NCAR in Broomfield, Colorado (40°N, 105°W), UIUC, Illinois (40°N, 88°W), and Goddard Space Flight Center (GSFC), Maryland (39°N, 78°W). The dots and lines for the measurements at MKO, NCAR, and GSFC indicate the average values and the ranges of the measured values, respectively. The dot for the UIUC data represents a regression fitted value for the middle of November. The regression fit was performed over the data obtained from 1984 to 1986. The error bar for the UIUC data indicates the rms difference between the regression fit and the measured values.

are quite comparable to those measured in Colorado, Illinois, and Maryland, and during the airborne observations. This appears to indicate that the magnitudes of horizontal winds over the Pacific Ocean are comparable to those over the land masses. However, it should be noted that the observations in Hawaii were made in January while all the other observations were obtained in October or November.

The major characteristics of the vertical and horizontal wavenumber spectra for both eastward and westward flights are summarized in Table 3.3. Because of Doppler effects, the slopes and amplitudes of the horizontal wavenumber spectra were quite different depending on flight directions. The changes of the spectra could be also related to the changes in the rms horizontal wind velocities; the spectral slopes were steeper when the wind velocities were smaller. The average slope of the horizontal wavenumber spectra was -1.25. The spectral slopes computed for only the bottomside of the layer were consistently steeper than those computed for the topside. This is illustrated in Figure 3.18 where the bottomside and topside spectra for the eastbound and westbound legs of the eastward flight, and the eastbound and westbound legs of the westward flights are plotted. Notice that the two quasi-monochromatic waves corresponding to spectral peaks A and C in Figure 3.5 are dominant only in the bottomside spectra plotted in Figure 3.18 a and b. The topside spectral amplitudes are larger which indicates that the wave amplitudes were growing with altitude. The shallower slope on the topside may be the result of saturation affecting the larger scale waves. The average slope for the bottomside spectra for all four flight legs was -1.51 and the average slope for the topside spectra was -0.97. The changes of the slopes and amplitudes of the vertical wavenumber spectra appear to be related to the rms horizontal wind velocities. For example, during the eastward flight the spectral slope was steeper, the spectral amplitude was smaller, and the velocities were smaller on the eastbound leg compared to the westbound leg. The average slope of the vertical wavenumber spectra for all flights was -2.67.

Table 3.3. Summary of the Horizontal and Vertical Wavenumber Spectra Parameters Measured on the Eastward and Westward Flights

Flight	Eastward		Westward	
Direction of Flight	Eastbound	Westbound	Eastbound	Westbound
Date	Nov 15	Nov 15-16	Nov 18	Nov 17-18
Time (MST)	2210-2349	2355-0255	0130-0429	2142-0116
Flight Duration (hours)	1.65	3.00	2.98	3.57
Flight Distance (km)	1062	1295	1933	1722
Aircraft Velocity ^a (m s ⁻¹)	181±18	121±18	181±19	135±20
Vertical Wavenumber Spectrum				
Slope ^b	-2.80	-2.68	-2.49	-2.70
Altitude Range (km)	82-101	82-102	81-102	83-99
Horizontal Wavenumber Spectrum				
Slope ^c	-1.41	-1.12	-1.03	-1.43
Altitude Range (km)	82-101	82-102	81-102	83-99
Bottomside Horizontal Wavenumber Spectrum				
Slope ^c	-1.66	-1.29	-1.41	-1.69
Altitude Range (km)	82-90	82-90	81-90	83-90
Topside Horizontal Wavenumber Spectrum				
Slope ^c	-1.09	-1.01	-0.71	-1.05
Altitude Range (km)	90-101	90-102	90-102	90-99

^aLongitudinal component of the ground velocity of the aircraft.

^bSpectral slopes computed over vertical scales ranging from 2 to 10 km.

^cSpectral slopes computed over horizontal scales ranging from 70 to 700 km.

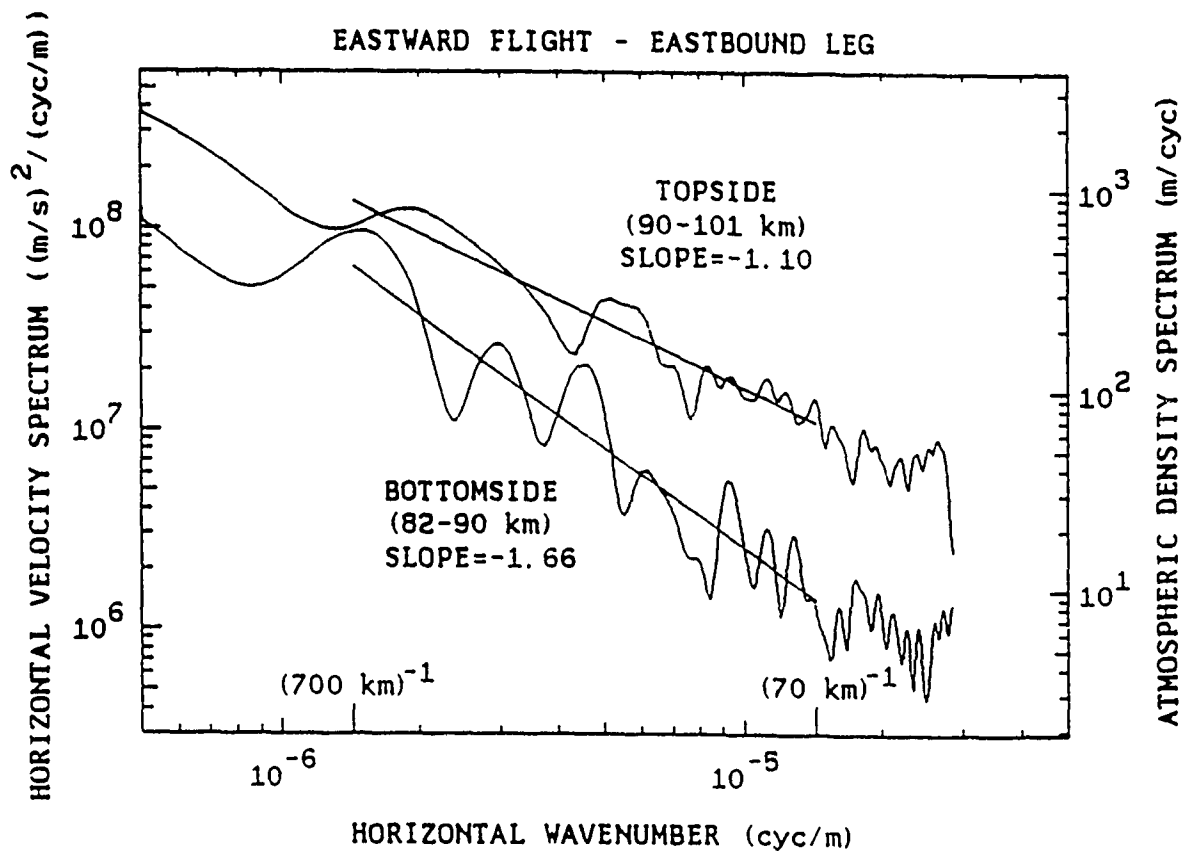


Figure 3.18 a). Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the eastbound leg of the eastward flight on November 15-16, 1986.

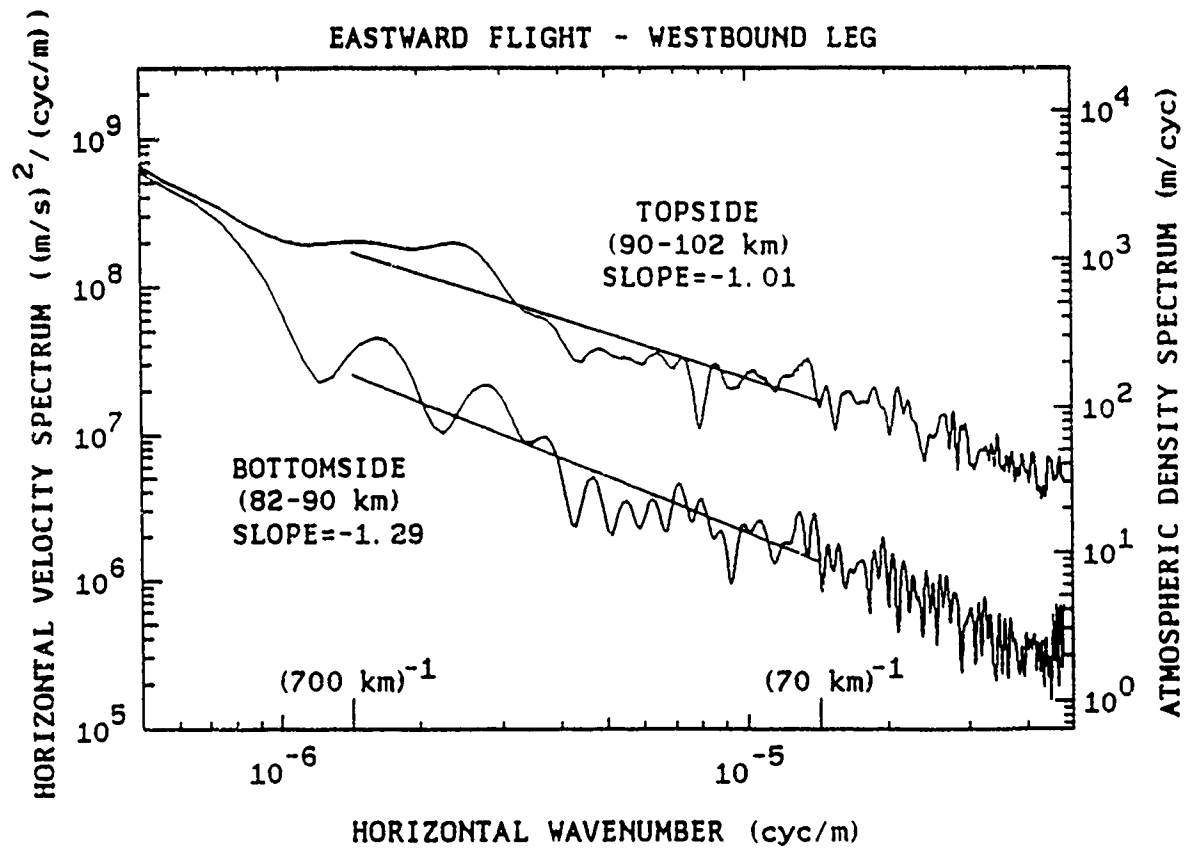


Figure 3.18 b). Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound leg of the eastward flight on November 15-16, 1986.

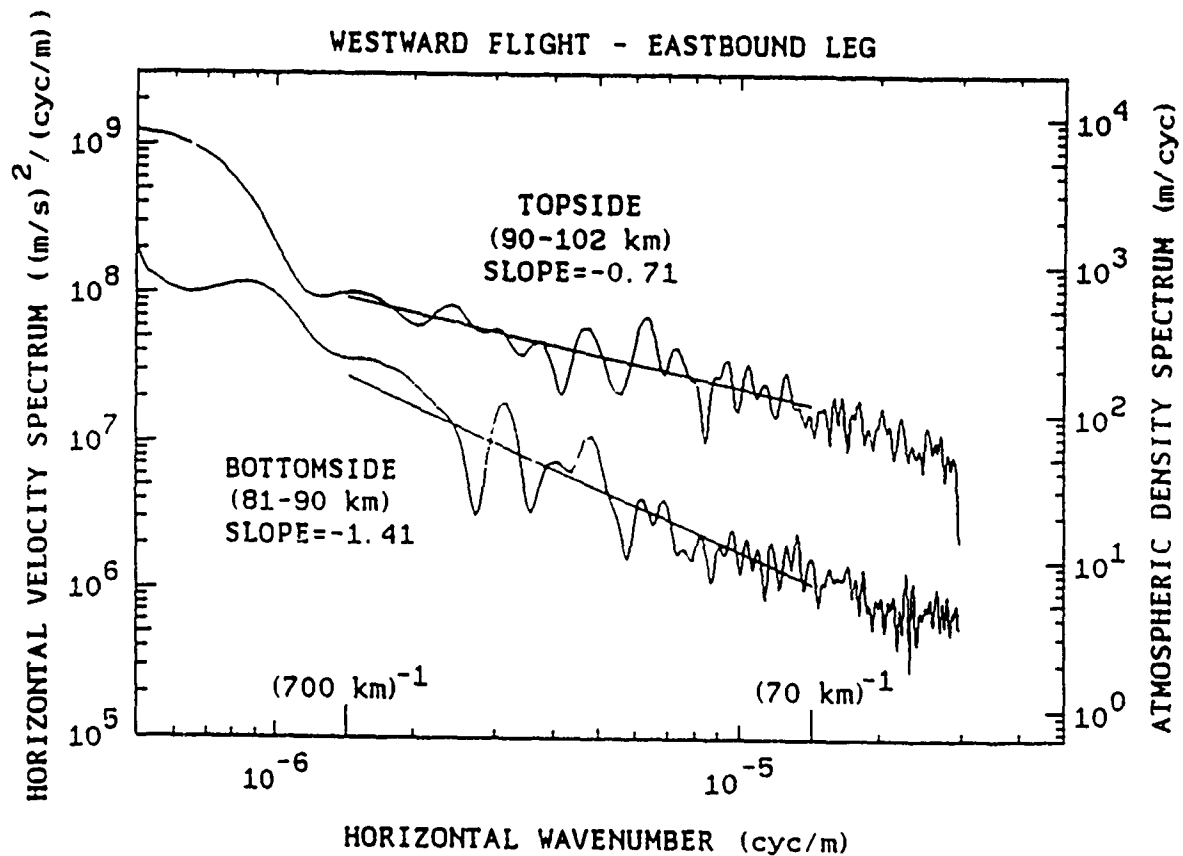


Figure 3.18 c). Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the eastbound leg of the westward flight on November 17-18, 1986.

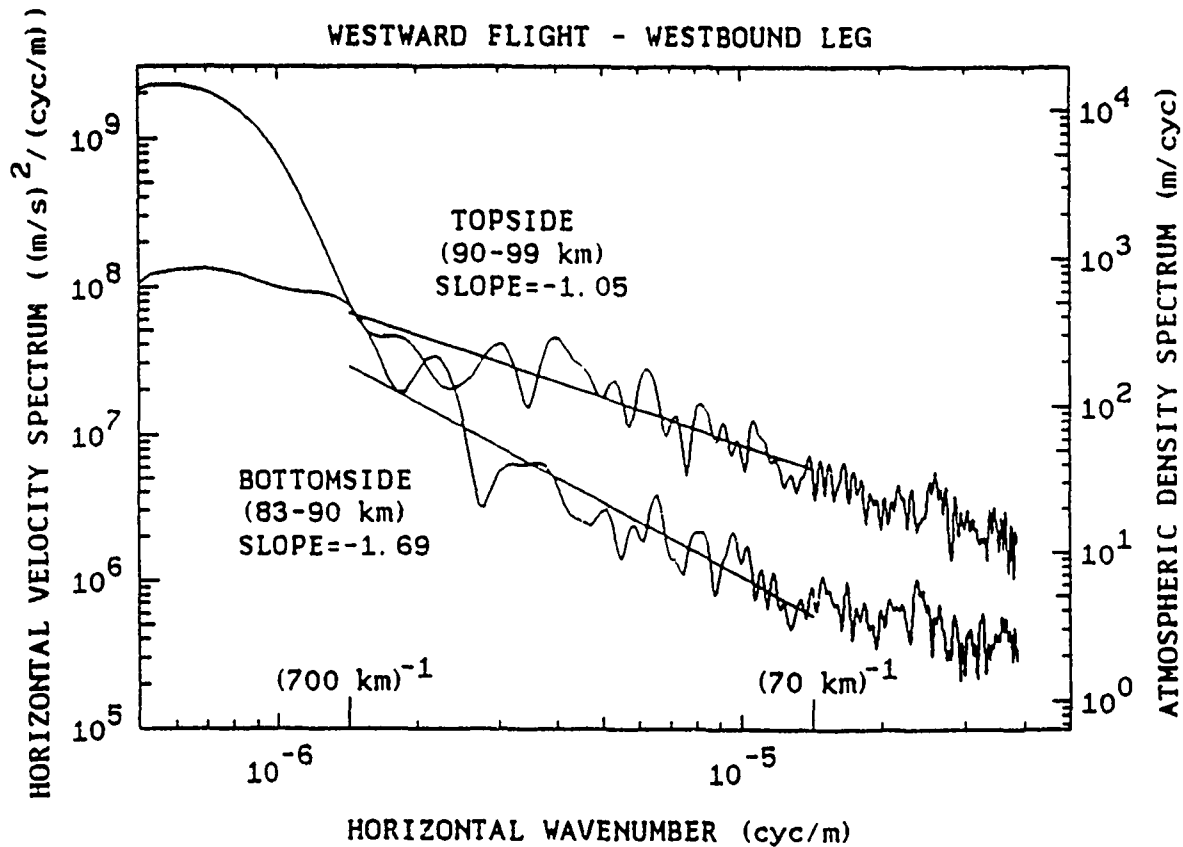


Figure 3.18 d). Horizontal wavenumber spectra computed for the bottomside and topside of the Na layer measured during the westbound leg of the westward flight on November 17-18, 1986.

3.5 Discussion

The mechanisms responsible for the observed density and velocity perturbations in the troposphere, stratosphere, and mesosphere have been the subject of much speculation. Two quite different mechanisms have been suggested to explain horizontal wavenumber spectra of density and velocity perturbations. The two mechanisms are quasi two-dimensional turbulence and internal gravity waves. Both mechanisms predict a $k_x^{-5/3}$ dependence in horizontal wavenumber spectra [Gage and Nastrom, 1985; 1986]. The quasi two-dimensional turbulence was suggested by Gage and Nastrom [1985; 1986] for the horizontal wavenumber spectra of meridional and zonal velocities obtained from the Global Atmospheric Sampling Program (GASP). The GASP data were collected with instrumentation on Boeing 747 aircraft at altitudes ranging from 9 to 14 km. The slopes of both zonal and meridional wavenumber spectra were approximately -3 at horizontal scales from 700 km to 10000 km, and -5/3 at scales from 3 km to 500 km. When Gage and Nastrom applied the Taylor transformation to temporal frequency spectra, they found that the transformed frequency spectra agreed with the GASP horizontal wavenumber spectra in both amplitude and shape. The temporal frequency spectra were computed from the data obtained at altitudes near 86 km with an ST/MST radar. Hence, Gage and Nastrom suggested that the quasi two-dimensional turbulence was responsible for the spectra.

Fritts *et al.* [1989] suggested that internal gravity waves were responsible for their horizontal wavenumber spectra. Their spectra were calculated from atmospheric density variations measured during seven re-entries of NASA space shuttles in the altitude range from 60 to 90 km. The spectra exhibited a slope of approximately -2 at horizontal scales from about 10 to 1000 km.

The horizontal wavenumber spectra inferred from the airborne Na lidar data appear to be the result of internal gravity waves instead of turbulence for the following reasons. First, the horizontal and vertical variations of the Na layer were apparently influenced by waves. For

example, during the eastward flight the variations of the centroid height and layer shape were significantly influenced by waves. Second, the horizontal wavenumber spectra computed from the lidar data exhibited distinct spectral peaks which appear to be characteristics of waves. Interestingly, the lidar observation of the wave with a period of 6.1 hours on the eastward flight seems to agree with radar observations. During the period from November 4 to 20, 1986, the data collected with the ST radar in Platteville, Colorado showed strong 6 hour variations in winds in the altitude range of the Na layer (see Section 4.3). Although only the zonal parameters of the 6.1-hour period wave were explored with the airborne Na lidar, the estimated wave parameters appear to be the intrinsic parameters. This wave seems responsible for the 6-hour period wind variations observed with the Platteville radar, and is probably tidally driven because of its period and westward phase propagation. Finally, Doppler-shifting effects which depended on flight directions were clearly apparent in the horizontal wavenumber spectra. As a consequence, the spectra changed significantly in amplitude and shape when flight directions changed. Some spectral peaks were Doppler shifted to lower wavenumbers, while others were shifted to higher wavenumbers.

From ground-based Na lidar observations of quasi-monochromatic gravity waves obtained at Urbana, Illinois, *Gardner and Voelz* [1987] estimated the slope of the wave kinetic energy distribution to be -2.95 for vertical scales from 2 km to 17 km, and -1.05 for horizontal scales from 20 km to 3000 km. These kinetic energy distribution slopes are comparable to the slopes of the vertical and horizontal wavenumber spectra calculated in this paper from the airborne Na lidar data. The average slope of the vertical wavenumber spectra of the airborne lidar data was -2.67 at scales from 2 km to 10 km, and the average slope of the horizontal wavenumber spectra was -1.25 at scales from 70 km to 700 km.

The slope of the horizontal wavenumber spectra calculated from the airborne lidar data appears to be shallower than the slopes of the GASP [*Gage and Nastrom*, 1985; 1986] and shuttle re-entry data [*Fritts et al.*, 1989]. The slope of the GASP data was about -3 at horizontal

scales from 700 km to 10000 km and $-5/3$ from 3 km to 500 km. The spectral slopes of the airborne lidar data do appear to increase significantly for horizontal scales greater than 700 km. However, because the distances of the flight legs varied from 1062 km to 1933 km, the spectral amplitudes at scales larger than 700 km should be viewed with caution. The differences between the GASP and airborne lidar data may stem from the difference in the altitude range of the two data sets. The altitude range of the airborne lidar data was from 82 to 101 km, and that of the GASP data was from 9 to 14 km. The horizontal spectra slopes of the airborne lidar data are also shallower than the slope of about -2 estimated from the shuttle re-entry data.

The slopes of the vertical wavenumber spectra calculated from the airborne lidar data appear to be comparable to the slopes obtained from radar observations. For example, *Smith et al.* [1985] calculated the spectral slopes of -2.0 to -2.8 from data collected with an MST radar at Poker Flat, Alaska in the altitude range from 82 to 88 km. The amplitudes of the vertical wavenumber spectra of the airborne lidar data also agree within a factor of 3 with those of the spectra observed with radar systems [*Tsuda et al.*, 1988; *Vincent*, 1984].

The horizontal velocity variances computed from the airborne lidar data also appear to be comparable to those calculated from the ground-based lidar data and the radar data, but larger than those calculated from the shuttle re-entry data. The horizontal velocity variances computed from the airborne lidar data ranged from 310 to $1800 \text{ m}^2 \text{ s}^{-2}$ with a mean of $1100 \text{ m}^2 \text{ s}^{-2}$. These variances are comparable to those computed from the ground-based lidar observations at Mauna Kea Observatory, Hawaii, at Broomfield, Colorado, at Urbana, Illinois, and at the Goddard Space Flight Center, Maryland. The variances of the radar data collected in Alaska ranged from 650 to $1300 \text{ m}^2 \text{ s}^{-2}$ [*Balsley and Carter*, 1982; *Balsley and Garelo*, 1985]. The variances of the shuttle re-entry data collected over the Pacific Ocean ranged from 55 to $415 \text{ m}^2 \text{ s}^{-2}$ [*Fritts et al.*, 1989].

3.6 Summary

The kinetic energy horizontal and vertical wavenumber spectra of horizontal winds are inferred from the airborne lidar measurements of Na density profiles. The airborne data were collected during two flights in November 1986. The two flights included one roundtrip from Denver, Colorado to Springfield, Illinois and another from Denver to the Pacific Coast. During these flights, the data were collected during the 11.2 hours of flights over baselines totaling 6000 km. The average slope of the horizontal wavenumber spectra was -1.25 at horizontal scales ranging from 70 to 700 km, and the average slope of the vertical wavenumber spectra was -2.67 at vertical scales from 2 to 10 km. The altitude range of the measurements was approximately 82 to 101 km. The slopes of the horizontal wavenumber spectra computed only for the bottomside (82-90 km) of the Na layer were consistently steeper than those computed for the topside (90-101 km). The average slope of the bottomside horizontal wavenumber spectra was -1.51, and that of the topside spectra was -0.97.

Internal gravity waves appear to be responsible for major features of the airborne Na lidar data. The observed features include the systematic horizontal and vertical variations of the Na density profiles, the horizontal variations of the centroid height, the presence of distinct spectral peaks in the horizontal wavenumber spectra, and Doppler-shifting of these spectral peaks. It is difficult to interpret these observed features in terms of turbulence. However, it seems possible that the dominant density perturbations observed in the lidar data at horizontal scales in the range from 20 to 2000 km are due to gravity waves.

The slopes of the vertical wavenumber spectra agree well with the slopes calculated from radar observations. The slopes of the horizontal wavenumber spectra of the airborne lidar data are shallower than the slopes of the GASP and shuttle re-entry spectra. In general, the rms horizontal wind velocities increased with time and with longitude from the Pacific Coast to the Great Plains. The rms horizontal wind velocities inferred from the airborne data are comparable with those inferred from ground-based lidar data obtained in Hawaii, Colorado, Illinois, and

Maryland. This appears to indicate that the rms horizontal wind velocities measured over the Pacific Ocean are comparable to those measured over the land masses.

4. CORRELATIVE RADAR AND AIRBORNE SODIUM LIDAR OBSERVATIONS OF THE VERTICAL AND HORIZONTAL STRUCTURE OF GRAVITY WAVES AND TIDES NEAR THE MESOPAUSE

4.1 Introduction

For the past several decades, radar techniques have made important contributions to the understanding of the dynamics of the mesopause region. In recent years, Na lidar techniques have matured to the extent that lidar studies of wave dynamics are now complementing radar observations. *Vincent and Fritts* [1987] recently reported a radar-based climatology study of gravity waves measured at Adelaide, Australia. *Gardner and Voelz* [1987] also reported extensive lidar-based observations of monochromatic gravity waves measured at Urbana, Illinois. Studies of atmospheric tides using Na lidars have been reported by *Batista et al.* [1985] and *Kwon et al.* [1987].

Joint observations of radar and Na lidar for exploring dynamics of the mesopause region have been rare. The only joint observations were reported by *Avery and Tetenbaum* [1983]. The measurements were conducted at Urbana, Illinois in January and March, 1980 using the Urbana meteor radar and Na lidar. The results showed some correlation between the altitude of maximum Na density and the altitude of the zero wind node of the prevailing wind. The reconstruction of the neutral atmosphere density perturbation profile from meteor wind harmonics exhibited small-scale fluctuations similar to those observed in the Na density profiles.

Because most radar and lidar observations have been made typically at fixed elevation angles, measurements of the horizontal structure of gravity waves have been limited. By using Doppler radars and dual bistatic radars, the horizontal structure of gravity waves was studied by *Vincent and Reid* [1983], *Smith and Fritts* [1983], and *Meek et al.* [1985]. By using multiple spaced lidars and steerable lidars, the horizontal structure of the Na layer was also studied by

Thomas et al. [1977], *Clemesha et al.* [1981], and *Gardner et al.* [1982; 1986]. Because of the increased atmospheric attenuation and longer propagation paths at the lower elevation angles, steerable lidar measurements are limited to a baseline of about 100 km or less.

To study longer baselines, the UIUC group conducted an airborne lidar experiment with the support of the NCAR-RAF in Broomfield, Colorado in November 1986. To investigate the longitudinal characteristics of gravity waves and tides at mid-latitudes, a total of three flights were made over the Great Plains, Rocky Mountains, and Pacific Coast. The flights were conducted out of Stapleton International Airport in Denver, Colorado using the NCAR Electra aircraft. In addition to the airborne observations, the lidar was operated on the ground in Broomfield and Denver during several nights from November 7 to 20. During the ground-based and airborne lidar observations, the University of Colorado group operated an ST radar in Platteville, Colorado. The results of these joint lidar/radar observations are presented in this chapter.

4.2 Description of the Experiment

4.2.1 Radar

From November 4 to 20, 1986, the ST radar located at Platteville, Colorado (40°N, 104°W) was operated by the University of Colorado group. In order to measure wind perturbations in the mesosphere and lower thermosphere, a meteor echo detection and collection system was attached to the ST radar, and operated. This system is described in detail by *Wang et al.* [1988]. Two radar beams were directed at azimuth angles of 0° (north) and 90° (east). The zenith angles of both beams were 15°, and the operating frequency was 49.8 MHz.

4.2.2 Lidar

In early November of 1986, the UIUC lidar was installed on board the Electra aircraft operated by the NCAR-RAF in Broomfield, Colorado. The system included a flashlamp-pumped dye laser, a 35 cm diameter Cassegrain telescope, and associated electronic and optical subsystems. The major lidar and aircraft parameters were summarized in Table 2.1. Following the installation, ground-based lidar observations were obtained in Broomfield for a total of 17.5 hours on 4 different nights from November 7 to 20. Additional ground-based observations were obtained in Denver as pre- and post-flight tests for a total of 2 hours on 3 different nights from November 13 to 18. The airborne lidar observations were described in detail in Section 3.3. Table 4.1 summarizes the system performance and observation times for the ground-based and airborne lidar observations.

4.3 Radar Observations

The vertical profiles of average meridional and zonal background wind velocities measured from November 4 to 20, 1986 are plotted in Figure 4.1 a and b. The average zonal background wind velocity was approximately 5 m s^{-1} eastward in the altitude range of the Na layer (80-105 km). The average meridional background wind velocity was small compared to the zonal background wind velocity at altitudes lower than 96 km. However, the meridional background wind velocity appears to increase at a higher altitude than 96 km.

The diurnal variation of the horizontal wind perturbations measured near 90 km is plotted in Figure 4.2. The boxes represent the hourly velocity measurements, and the solid curve represents a sinusoidal least-squares fit. The area of the boxes is proportional to the number of echoes used in computing the hourly average. The period of the dominant oscillation is approximately 6 hours, and the amplitude of the 6-hour period oscillation is about 1.3 m s^{-1} .

The vertical profiles of the 6-hour period meridional and zonal wind velocities are plotted in Figure 4.3. Note that the amplitudes of the zonal winds are much larger than the amplitudes of

**Table 4.1. Summary of the Airborne and Ground-based
Na Lidar Observations**

<u>Date</u>	<u>Time(MST)^a</u>	<u>Signal Level^b</u>		<u>Location</u>	<u>Comments</u>
		(count/shot)			
November 7-8	2259-0331	0.6		Broomfield	Ground-based
November 11	0236-0300	0.8		Broomfield	Ground-based
November 12-13	2019-0043	2		Broomfield	Ground-based
November 13	2015-2033	12		Denver	Ground-based
November 13	2119-2357	10		CO,NM,AZ,VY	Triangular flight
November 15	1943-2055	6		Denver	Ground-based
November 15-16	2210-0309	7		CO-IL	Eastward flight
November 15	0352-0357	4		Denver	Ground-based
November 17	2011-2033	6		Denver	Ground-based
November 17-18	2141-0502	8		CO-Pacific	Westward flight
November 19-20	2228-0625	6		Broomfield	Ground-based

Total Observation time

Ground-based: 19 hours 28 min

Airborne: 14 hours 58 min

^aMST = UT - 7 hours.

^bTypical signal level at Urbana is 5 counts/shot.

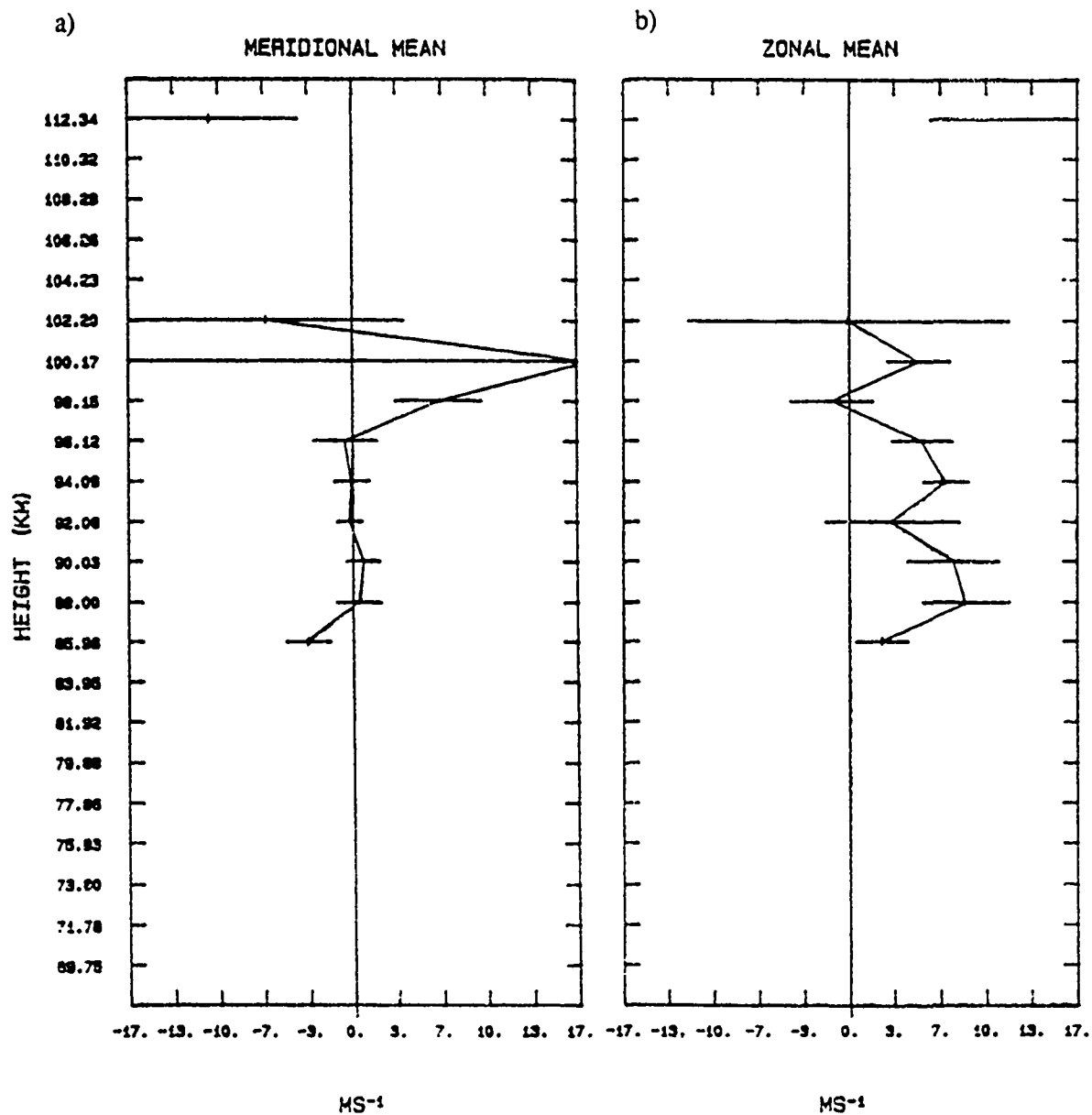


Figure 4.1. Vertical profiles of a) the meridional component and b) the zonal component of average background wind velocity measured with Platteville ST radar during the period from November 4 to 20, 1986. The data were provided by the University of Colorado.

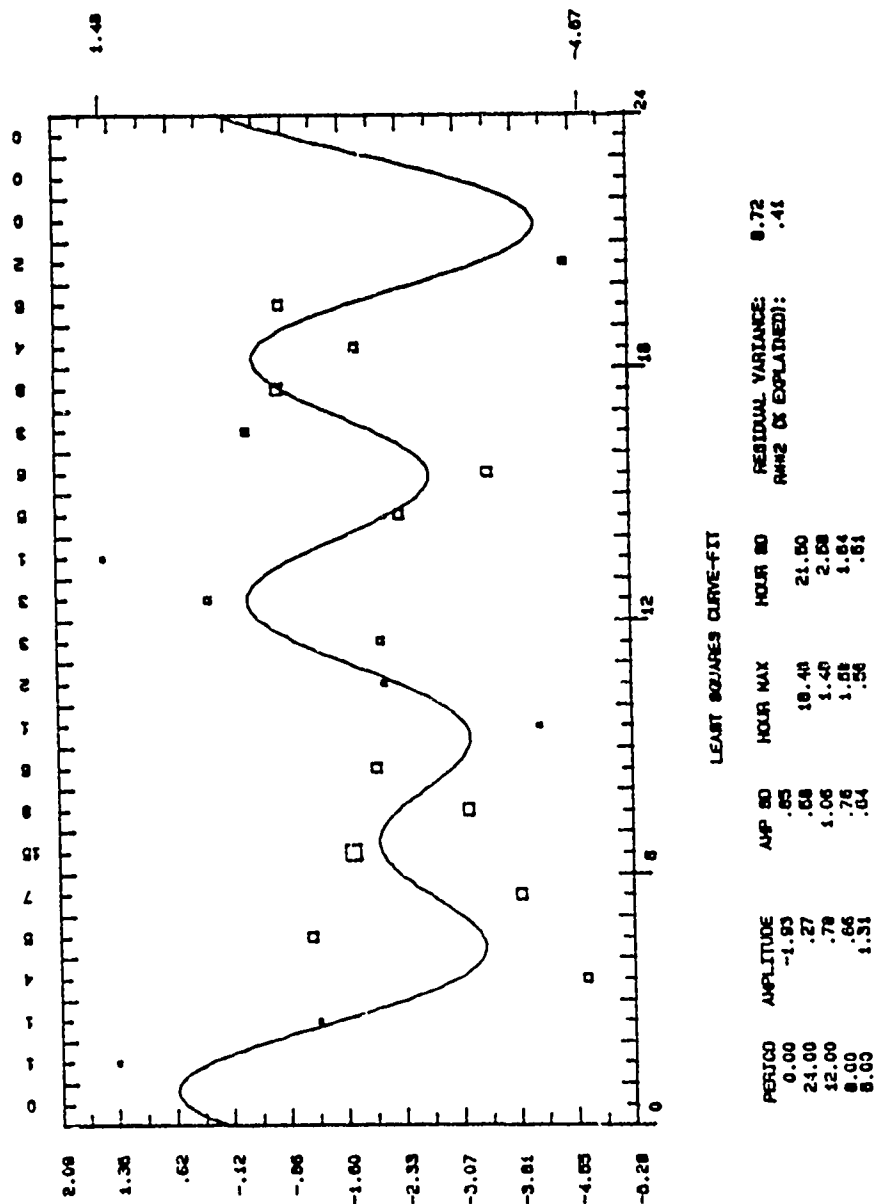


Figure 4.2. Diurnal variation of the horizontal wind perturbations measured with the Platteville ST radar during the period from November 4 to 20, 1986. The boxes represent the hourly velocity measurements, and the solid curve represents a sinusoidal least-square fit. The data were provided by the University of Colorado.

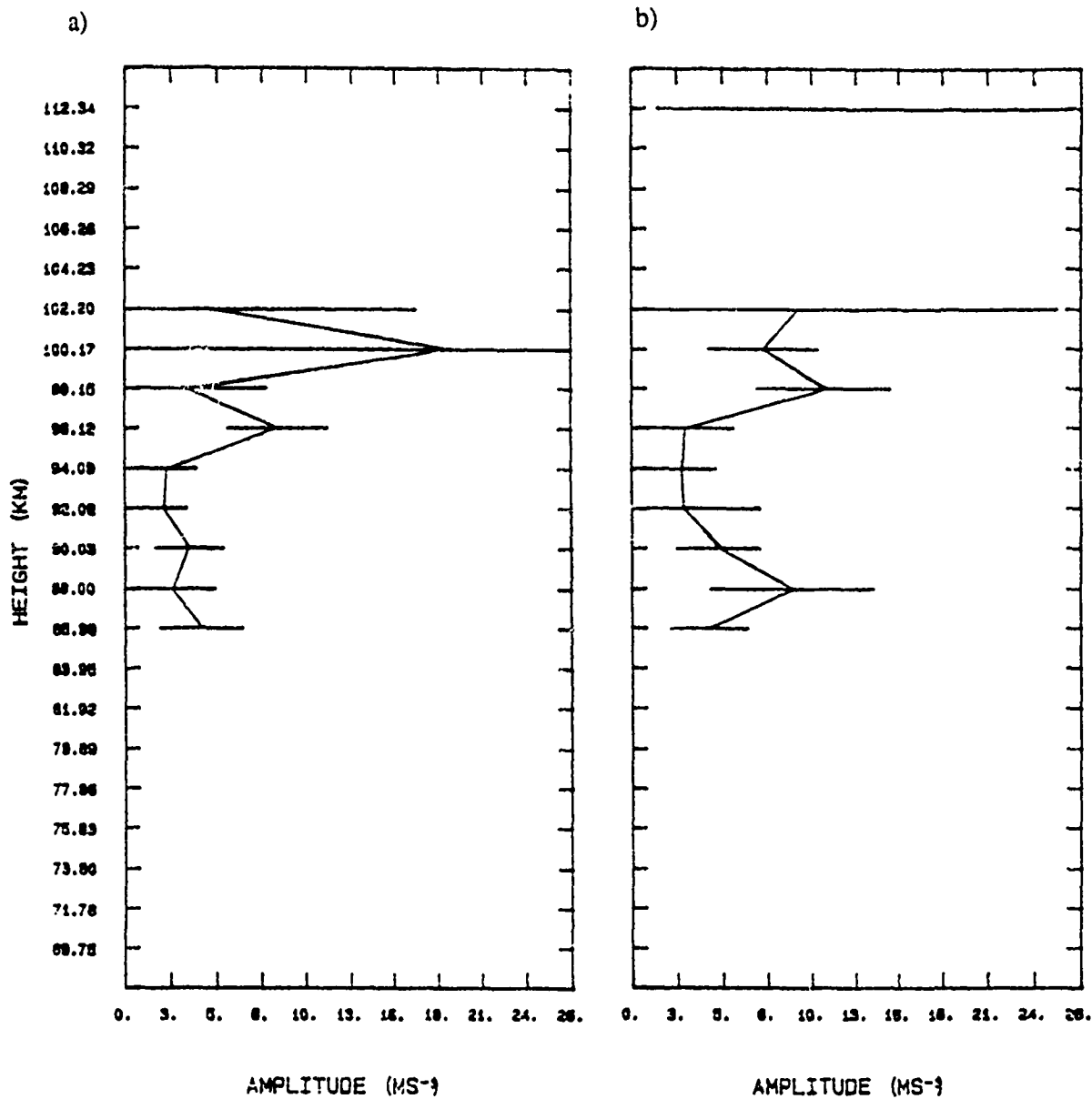


Figure 4.3. Vertical profiles of a) the meridional component and b) the zonal component of the 6-hour period horizontal winds. The data were provide by the University of Colorado.

the meridional winds in the altitude region from 85 to 92 km. This altitude region approximately corresponds to the bottomside of the Na layer. The largest zonal wind velocity in this altitude region was 9 m s^{-1} at 88 km. Large variations with periods of 6 hours were also observed with the ground-based Na lidar (Section 4.4) and the airborne Na lidar (Section 4.5), and were dominant only on the bottomside of the Na layer. These 6-hour period variations appear to be related to the 6-hour period variations in the horizontal winds measured with the radar.

4.4 Ground-based Lidar Observations

From November 7 to 20, 1986, a total of 19.5 hours of ground-based Na data were collected in Broomfield and Denver, Colorado with the lidar on board the NCAR Electra. In this section, the longest data set obtained on the night of November 19-20 in Broomfield will be presented. The time sequence of the density profiles measured on this night is plotted in Figure 4.4. The density profiles have been filtered vertically with a cutoff of 3 km and temporally with a cutoff of 50 min. The profiles have been normalized so that each has the same column abundance and are plotted on a linear scale at 10 min intervals.

The temporal variations of the layer centroid height, rms width, and column abundance are plotted in Figure 4.5. All three parameters show large variations with a period of approximately 6 hours. For example, the local maxima of the centroid height occurred at 2300 and 0452 MST. The time difference between these maxima was 5.9 hours. The periods of the dominant rms width and column abundance variations were approximately 6.6 hours and 7.0 hours, respectively. All three parameters also exhibit shorter temporal variations from 0000 to 0230 MST with a period of approximately 2 hours. For example, the local minima of the centroid height occurred at 0012 and 0155 MST for a time difference of 1.7 hours. The periods of the shorter temporal variations in the rms width and column abundance were 1.9 and 1.8 hours, respectively.

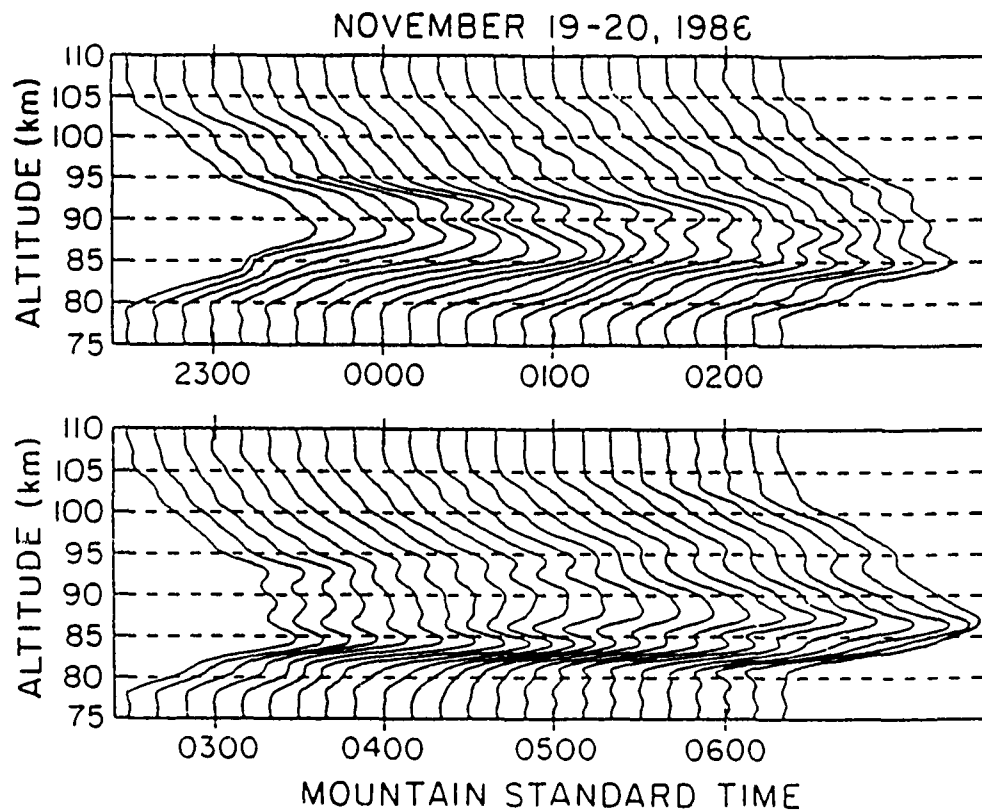


Figure 4.4. Sodium density profiles measured on the night of November 19-20, 1986. The profiles have been filtered vertically with a cutoff of 3 km and temporally with a cutoff of 50 min. The profiles have been normalized so that each has the same column abundance, and are plotted on a linear scale at 10 min intervals.

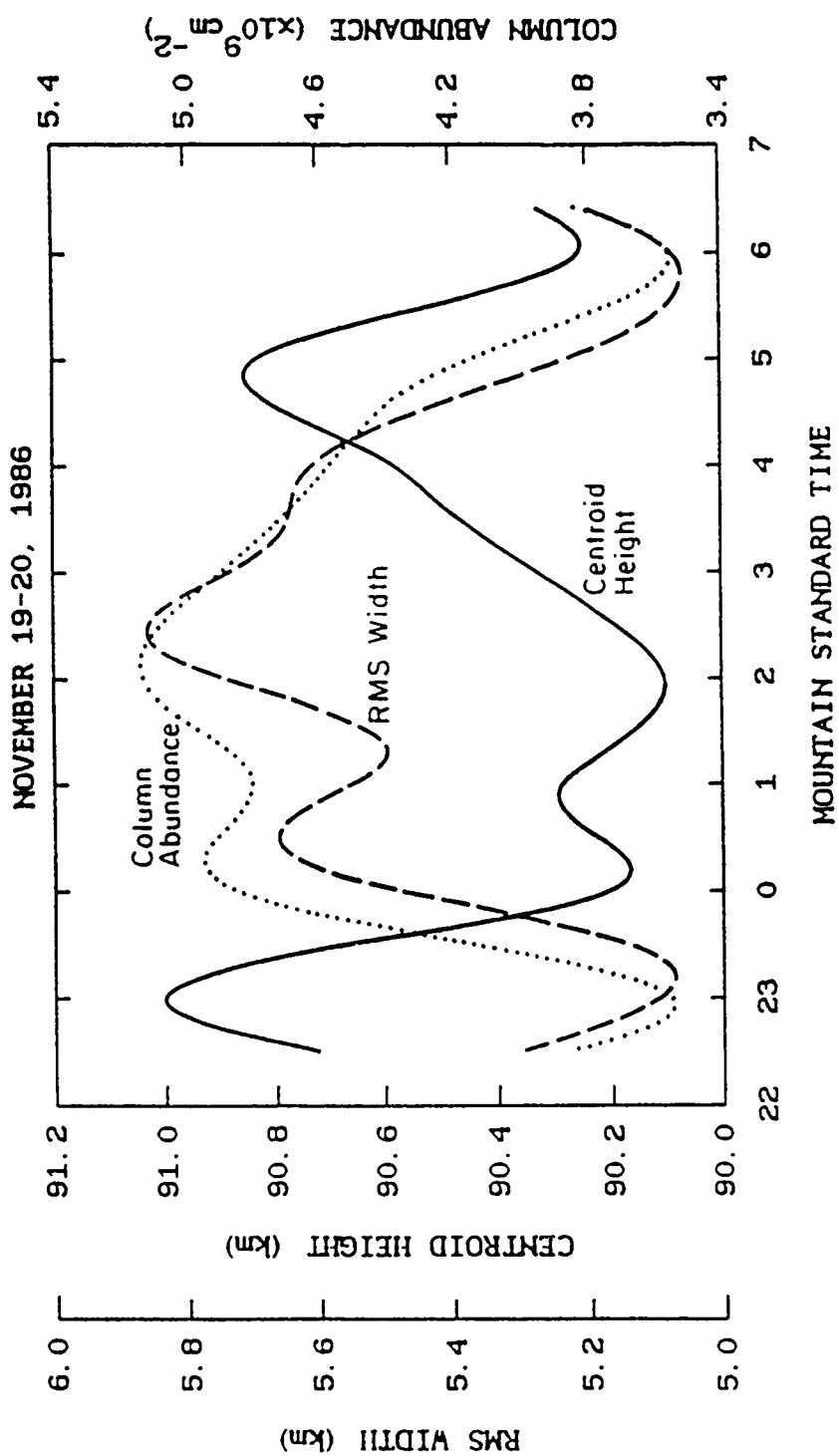


Figure 4.5. Temporal variations of the layer centroid height, rms width, and column abundance measured on November 19-20, 1986.

The shorter temporal variations were more dominant on the bottomside of the Na layer. The temporal frequency spectrum of the relative sodium density perturbations for the whole layer is plotted in Figure 4.6 a, and the temporal frequency spectra computed for the bottomside and the topside of the layer are plotted in Figure 4.6 b. The technique for computing the temporal spectra is described by *Gardner and Senft* [1989]. The Na density profiles were vertically filtered with a cutoff of 1 km before the spectra were calculated. The altitude range for the bottomside was 82 to 90 km, and that for the topside was 90 to 100 km. The straight lines in Figure 4.6 are linear regression fits which were used to estimate the spectral slopes over temporal scales from 0.5 to 6 hours. The spectral slopes were approximately -1.52 for the whole layer spectrum, -1.77 for the bottomside spectrum, and -1.12 for the topside spectrum. The topside spectral amplitudes are larger which indicates that the wave amplitudes were growing with altitude. The shallower slope of the topside spectrum may be the result of saturation affecting longer period waves. Note the dominant spectral peak in the bottomside spectrum near $1.5 \times 10^{-4} \text{ s}^{-1}$. The period is calculated to be 1.9 hours, which is quite comparable to the periods of the shorter temporal variations in the centroid height, rms width, and column abundance. The topside spectrum does not exhibit a spectral peak at the same period.

The shorter period oscillations are also seen clearly in temporal variations of Na density. Figure 4.7 shows the relative temporal variations from 81 to 92 km. The Na density profiles were first filtered vertically with a cutoff of 5 km and temporally with a cutoff of 90 min. The density variations were then computed and normalized so that the peak-to-peak amplitudes at each altitude were equal. The vertical phase progression of the shorter period variations is most clearly seen between 2230 and 0230 MST in the altitude range from 81 to 89 km. At higher altitudes, the variations with a period near 2 hours are not in evidence. The diagonal lines indicate the apparent phase progressions from which the vertical phase velocity is estimated to be 1.8 m s^{-1} . This velocity corresponds to a vertical wavelength of 12 km for a wave with the

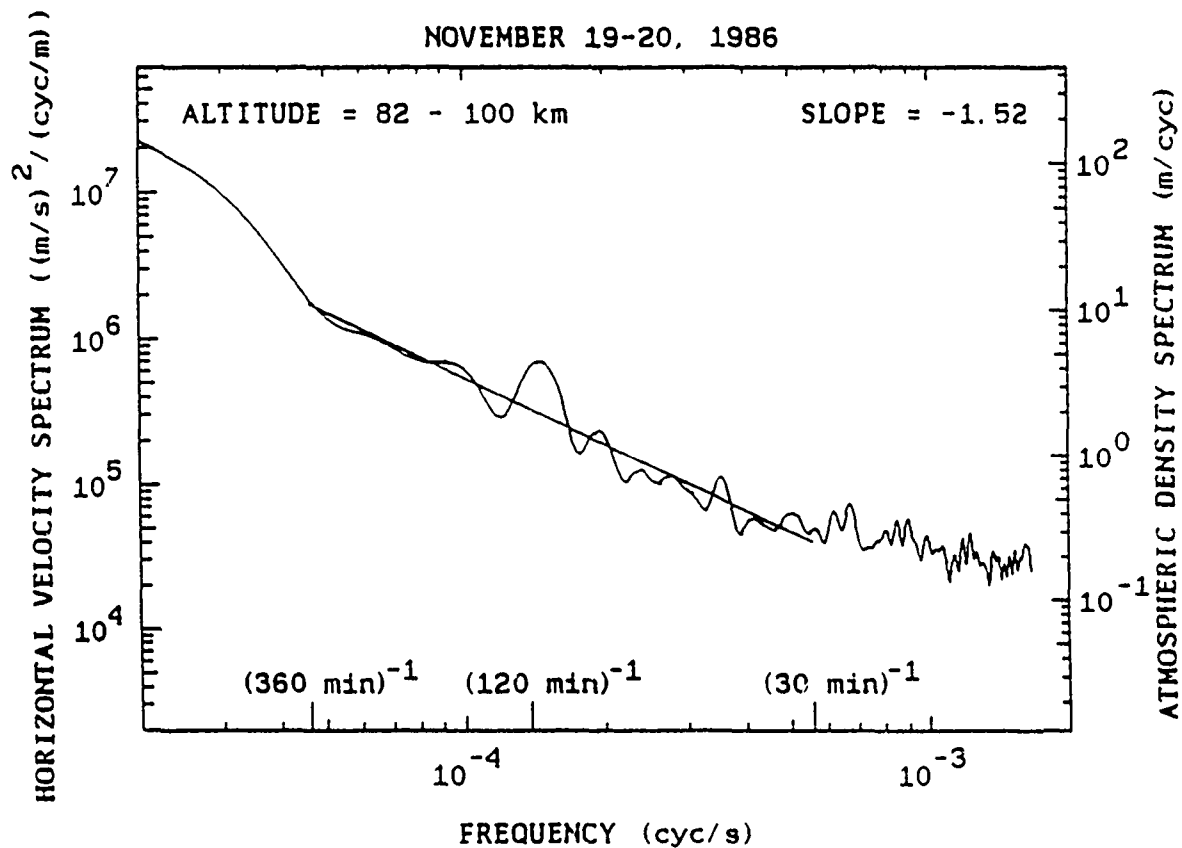


Figure 4.6 a). Temporal frequency spectrum computed for the whole layer from data collected on November 19-20, 1986. The Na density profiles were vertically filtered with a cutoff of 1 km before the spectra were computed. The straight line is a linear regression fit which was used to estimate the spectral slopes over temporal scales from 30 to 360 min.

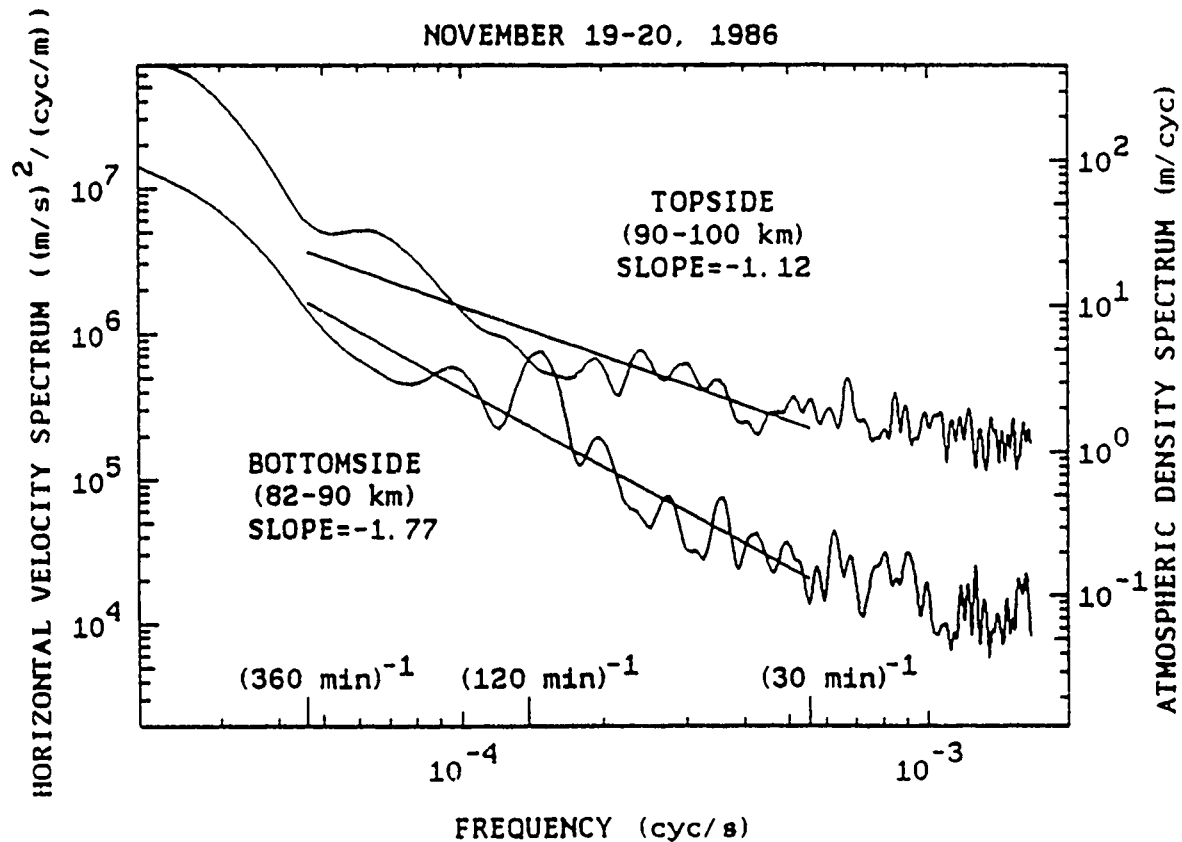


Figure 4.6 b). Temporal frequency spectra computed for the layer bottomside and topside from data collected on November 19-20, 1986. The Na density profiles were vertically filtered with a cutoff of 1 km before the spectra were computed. The straight lines are linear regression fits which were used to estimate the spectral slopes over temporal scales from 30 to 360 min.

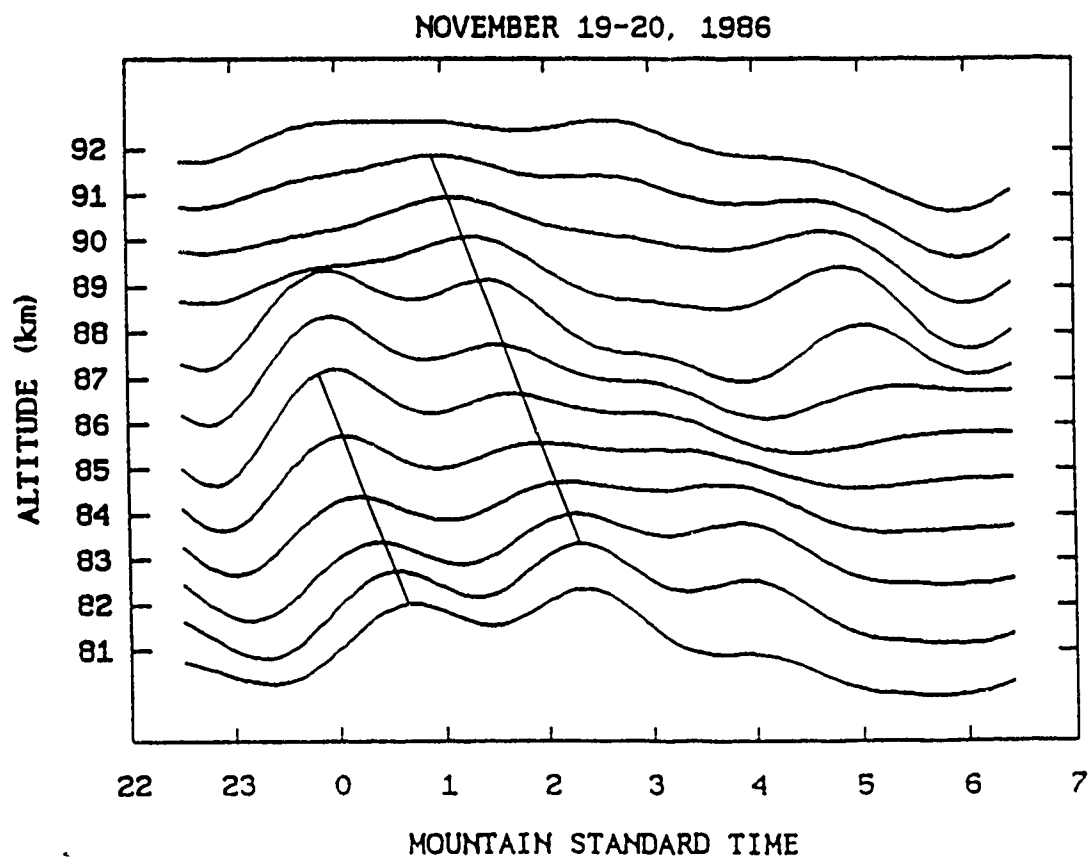


Figure 4.7. Relative temporal variations of the Na density from 81 to 92 km measured on November 19-20, 1986. The Na density profiles were vertically filtered with a cutoff of 5 km and temporally with a cutoff of 90 min. The diagonal lines indicate apparent vertical phase progressions. The estimated phase velocity is 1.8 m s^{-1} .

period of 1.9 hours. Because this wave may be Doppler shifted by the background winds, the intrinsic period and vertical wavelength may be different than these values.

The vertical wind velocity on the layer bottomside at 82 km is plotted versus time in Figure 4.8. The technique for estimating the vertical wind velocity is described in detail by *Kwon et al.* [1987]. To compute the vertical winds, the Na density profiles were filtered vertically with a cutoff of 5 km and temporally with a cutoff of 60 min. The period of the most dominant variation was 1.9 hours, and the amplitude of this variation decreased significantly with time toward the early morning of November 20. The vertical wind velocity on the layer topside at 101 km is plotted in Figure 4.9. Although a variation with the period of approximately 2 hours is present at this altitude, the amplitude of the 2-hour period oscillation is much smaller than at 82 km.

The average vertical wavenumber power spectrum of the Na density profiles obtained from 2300 to 0100 MST is plotted in Figure 4.10. This time interval corresponds to the period during which the amplitude of the 2-hour period wave was large. The dashed line in the spectrum indicates the estimated shot noise level. The technique for estimating vertical wavelengths of gravity waves from the Na profile vertical spectra is described in detail by *Gardner and Voelz* [1987]. The wavelengths of dominant waves in the vertical power spectrum in Figure 4.10 appear to be 6.4 km and 3.1 km. These wavelengths may represent harmonics of the vertical wavelength of the 2-hour period wave, which is about 12 km.

The rms horizontal wind velocity computed for the entire altitude range of the Na layer is plotted versus time in Figure 4.11. The technique for computing the rms horizontal winds is described by *Kwon et al.* [1989a]. The Na density profiles were initially filtered vertically with a cutoff of 1 km and temporally with a cutoff of 20 min. The wind amplitudes exhibit strong 2-hour period oscillations from 2300 to 0200 MST, and agree qualitatively with the temporal variations of the vertical winds estimated at the altitude of 82 km (Figure 4.8). The amplitude of the horizontal winds for the 2-hour period oscillations is about 7 m s^{-1} near midnight. The

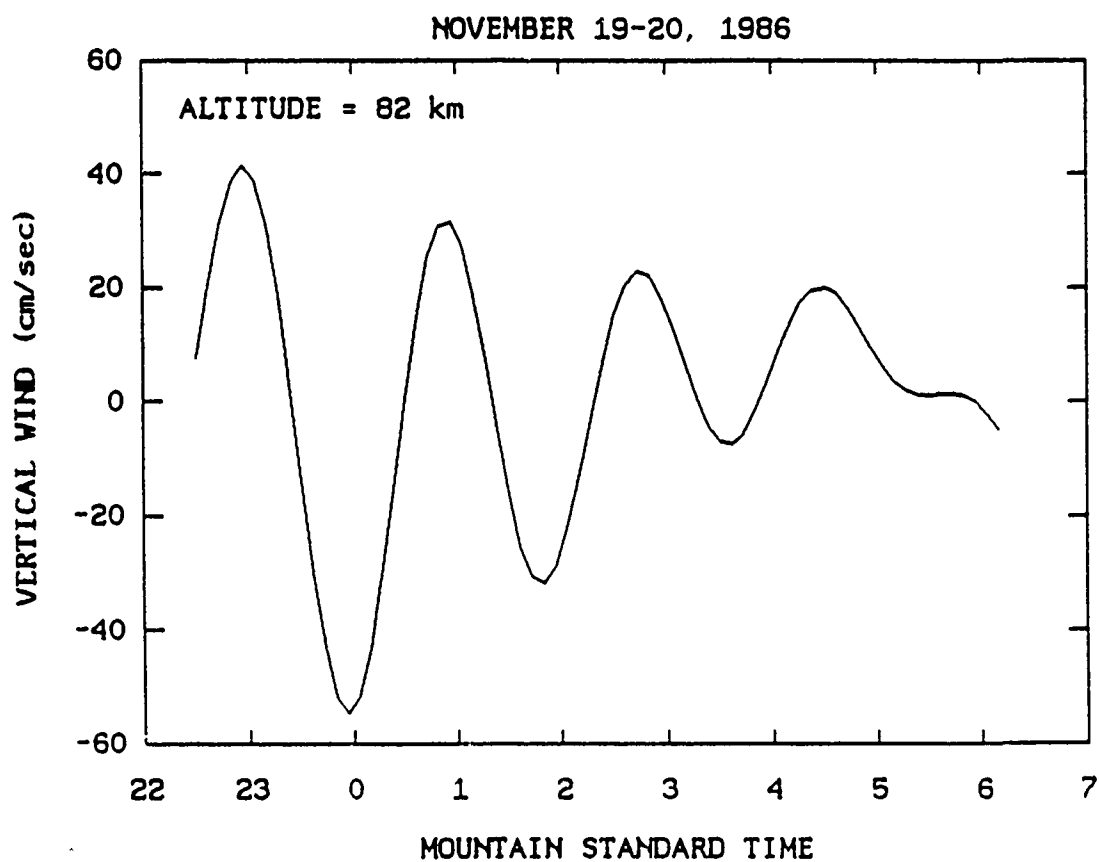


Figure 4.8. The vertical wind velocity estimated at the altitude of 82 km. The Na density profiles were initially filtered vertically with a cutoff of 5 km and temporally with a cutoff of 60 min.

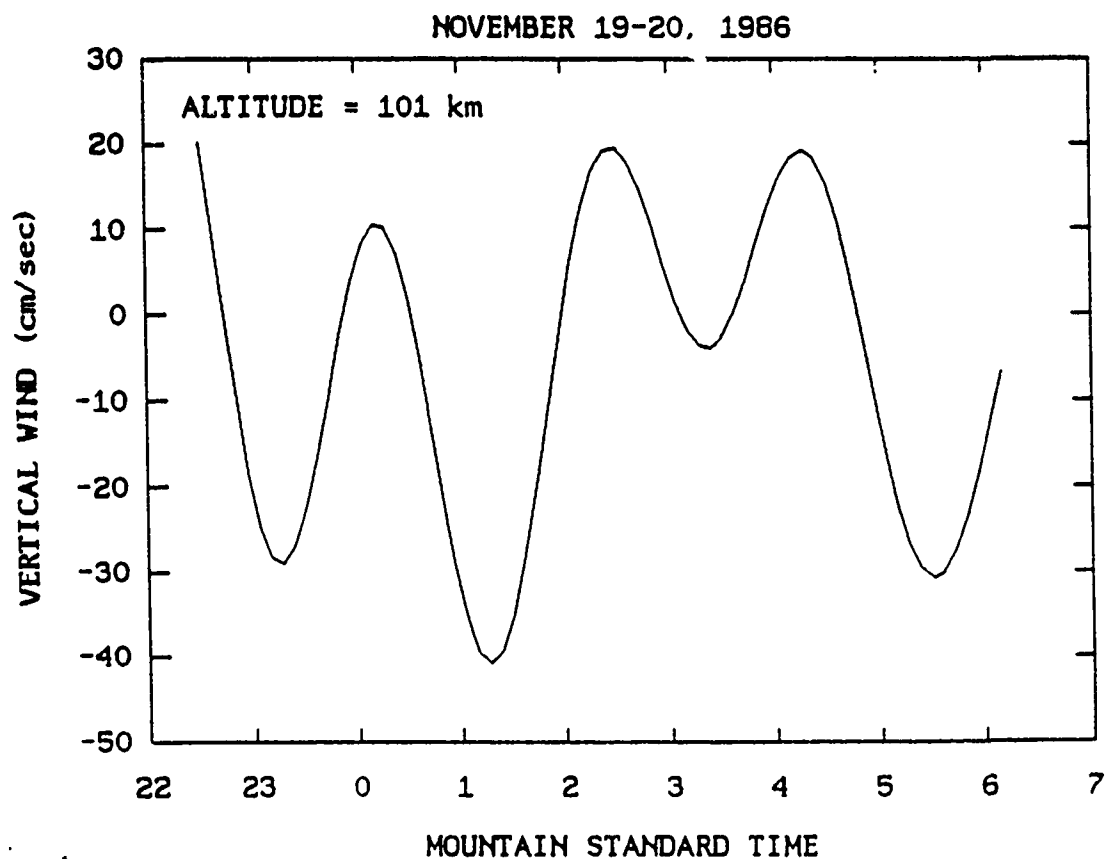


Figure 4.9. The vertical wind velocity estimated at the altitude of 101 km. The Na density profiles were initially filtered vertically with a cutoff of 5 km and temporally with a cutoff of 60 min.

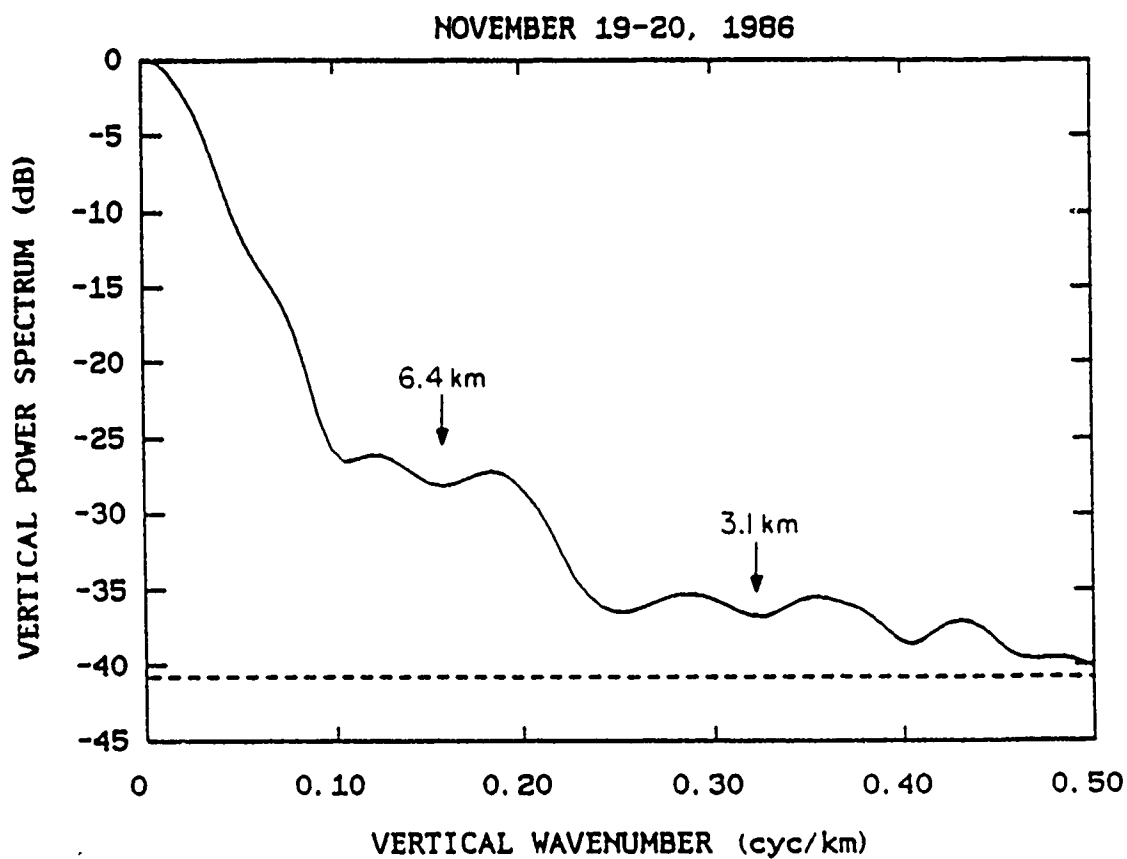


Figure 4.10. The average vertical wavenumber power spectrum computed for the Na density profiles collected from 2300 to 0100 MST on November 19-20, 1986.

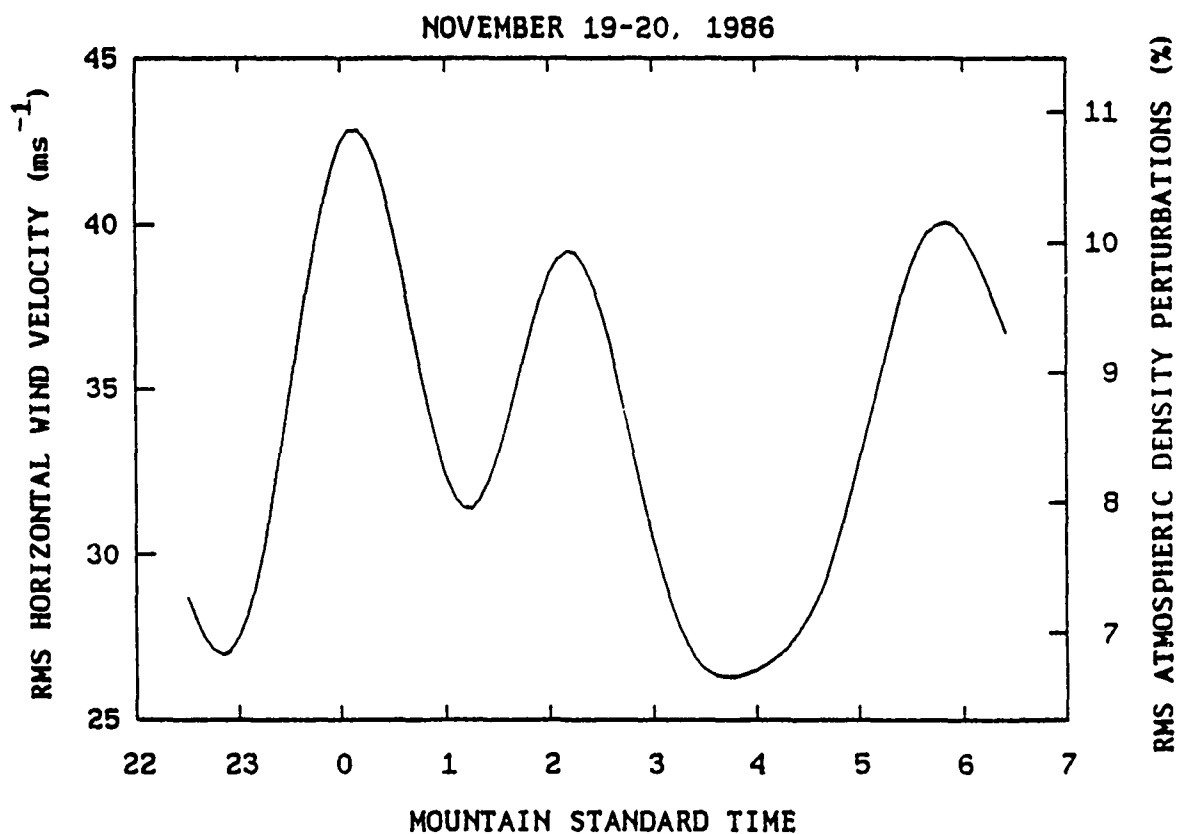


Figure 4.11. The rms horizontal wind velocity inferred from the Na density profiles collected on November 19-20, 1986. The profiles were initially filtered vertically with a cutoff of 1 km and temporally with a cutoff of 20 min.

horizontal wind amplitude of the 2-hour period wave is related to the amplitude of the spectral peak at the period of about 120 min in the temporal frequency spectrum (Figure 4.6 a) by

$$E_x(\omega_0) = \frac{\Delta T}{2} V_x^2(z_0) \left[\cosh(\beta \Delta z) - \frac{2\sigma_0^2}{\gamma H \Delta z} \sinh(\beta \Delta z) \right] \quad (4.1)$$

where $E_x(\omega_0)$ = spectral amplitude at frequency ω_0 ($\text{m}^2 \text{s}^{-1}$),

ΔT = observation period (s),

$V_x(z_0)$ = horizontal wind amplitude at the altitude of the centroid height of the Na layer (m s^{-1}),

β = amplitude growth factor (m^{-1}),

Δz = vertical extent of the Na layer (m),

σ_0 = rms width of the Na layer (m),

γ = ratio of specific heats (≈ 1.4), and

H = atmospheric scale height (m).

By assuming that $\beta \Delta z < 1$ for the 2-hour period wave, the horizontal wind amplitude which corresponds to the 2-hour spectral peak in Figure 4.6 a is calculated to be approximately 7 m s^{-1} . This velocity is for the horizontal winds averaged over the total observation period on November 19-20, and compares favorably with the velocity of 7 m s^{-1} estimated from the rms horizontal winds plotted in Figure 4.11.

For an atmospheric wind field generated by a monochromatic wave, *Hines* [1960] showed that the horizontal wind amplitude is related to the vertical wind amplitude by

$$\frac{V_x}{V_z} \approx \frac{T_{in}}{T_B} \quad (4.2)$$

where V_z = vertical wind amplitude,

T_{in} = intrinsic wave period,

T_B = Brunt-Vaisala period (≈ 5 min).

When the 2-hour period wave was strong near midnight, the amplitude of the horizontal winds was 7 m s^{-1} and that of the vertical winds was 0.3 m s^{-1} , which is the average of the amplitudes estimated at 82 km (Figure 4.8) and at 101 km (Figure 4.9). Thus $V_x/V_z = 23$ compares favorably with the ratio $T/T_B = 24$ for the 2-hour period wave.

Dominant waves with the period of 2 hours have been observed often at Urbana, Illinois [Gardner *et al.*, 1986; Gardner and Voelz, 1987; Gardner, 1989]. For example, the density profiles obtained during the early morning of June 12, 1984 [Gardner *et al.*, 1986] and June 28, 1988 [Gardner, 1989] exhibited large vertical displacements of the bottomside of the layer with a period of about 2 hours. During these two nights, the 2-hour period oscillation was dominant only on the bottomside of the layer. However, the vertical wavelength of 12 km measured on November 19-20 is larger than the vertical wavelengths of 2-hour period gravity waves which have been frequently observed at Urbana. The vertical wavelengths for the majority of the 2-hour period gravity waves ranged from 4.4 to 7.3 km [Gardner and Voelz, 1987].

The 6-hour period variations observed in the centroid height, rms width, and column abundance are also seen in the vertical variations of local Na density maxima plotted in Figure 4.12. The Na density profiles were initially filtered vertically with a cutoff of 3 km and then temporally with a cutoff of 50 min. There appear to be two maxima with dominant downward motions. Linear regression fits were used to estimate the vertical velocities. The upper density maximum moved downward at a velocity of 19.1 cm s^{-1} , and the lower density maximum moved at 18.3 cm s^{-1} . The average vertical velocity was 18.7 cm s^{-1} . The time separation between these maxima was 5.6 hours. The measured velocity corresponds to a vertical wavelength of about 4 km for a wave with the period of 6 hours. This vertical wavelength appears to be far too small to induce the large 6-hour period variations in the centroid height, rms width, and column abundance in the Na layer. Gardner and Shelton [1985] derived theoretical expressions for the layer parameter variations in terms of the wave period, vertical wavelength, and amplitude. Waves with vertical wavelengths less than 5 km have a negligible effect on the

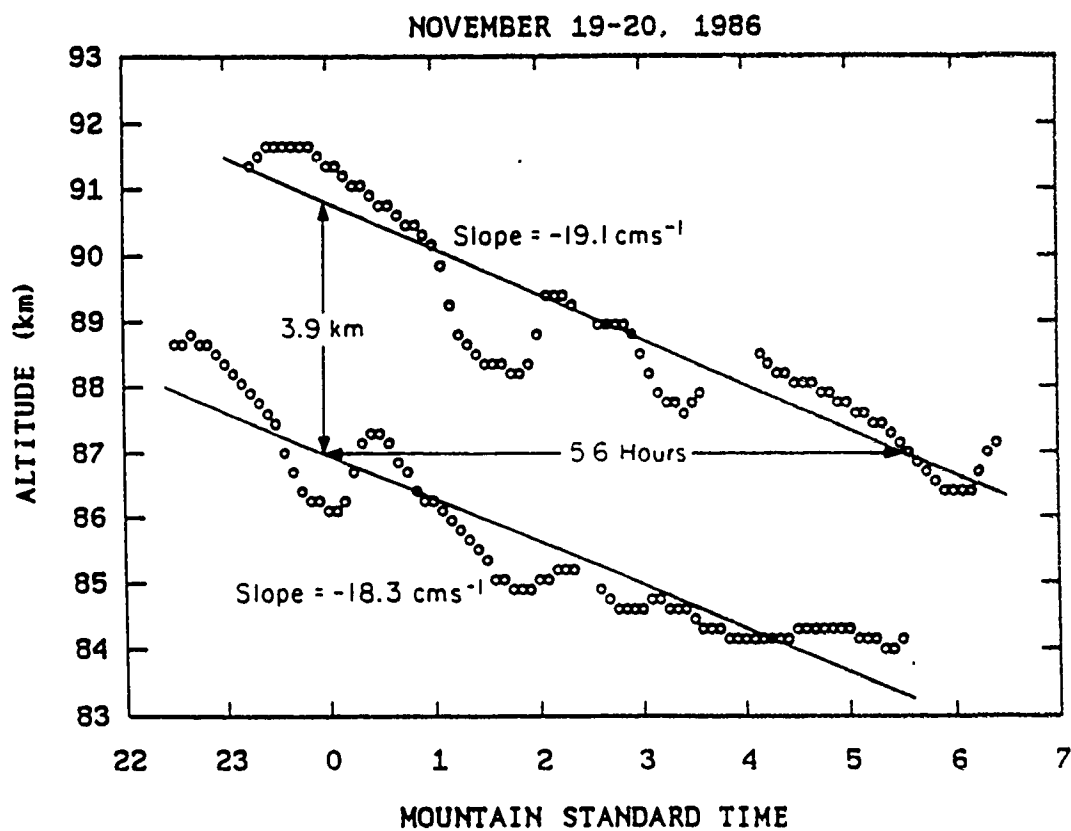


Figure 4.12. Temporal variations of the altitudes of local Na density maxima measured on November 19-20, 1986. The Na density profiles were initially filtered vertically with a cutoff of 3 km and temporally with a cutoff of 50 min.

layer centroid height, rms width, and column abundance. Appreciable perturbations in these parameters occur only for waves with vertical wavelengths exceeding about 10 km. Therefore, the vertical wavelength of the 6-hour period wave is believed to be much larger than 3.8 km.

To estimate the vertical wavelength of the 6-hour period wave accurately, the average vertical power spectrum for the Na density profiles obtained from 0300 to 0625 MST is computed and plotted in Figure 4.13. The vertical wavelength of a dominant wave in the spectrum is approximately 7.3 km, which may be the vertical wavelength of the 6-hour period wave. Although this vertical wavelength is larger than that estimated from the apparent vertical progressions of the Na density maxima, it is comparable to the vertical wavelength of the 6-hour period waves observed at Urbana, which ranged from 8.1 to 10.2 km [Gardner and Voelz, 1987]. It is also closer to the value estimated from the airborne observations [Kwon *et al.*, 1989a].

Gardner and Voelz [1987] described the technique for estimating the amplitude of the wave-induced atmospheric density perturbations from the Na profile vertical power spectra. From Figure 4.13, the amplitude of the atmospheric density perturbations due to the 6-hour period wave is computed to be 3.9 %. Gardner and Voelz [1987] also showed that this amplitude is related to the horizontal wind amplitude by the following equation.

$$V_x = \sqrt{\frac{\gamma H g}{\gamma - 1}} A e^{\beta z} \quad (4.3)$$

where g = gravitational acceleration ($= 9.5 \text{ m s}^{-2}$), and

$A e^{\beta z}$ = amplitude of the atmospheric density perturbations due to the wave.

The horizontal wind amplitude for the 6-hour period wave is computed to be 17 m s^{-1} .

The average vertical wavenumber spectrum corresponding to all the density profiles obtained on November 19-20 is plotted in Figure 4.14. The technique for estimating the vertical wavenumber spectrum is described by Kwon *et al.* [1989a]. The Na density profiles were first filtered temporally with a cutoff of 20 min, and then the spectrum was calculated. The shot noise

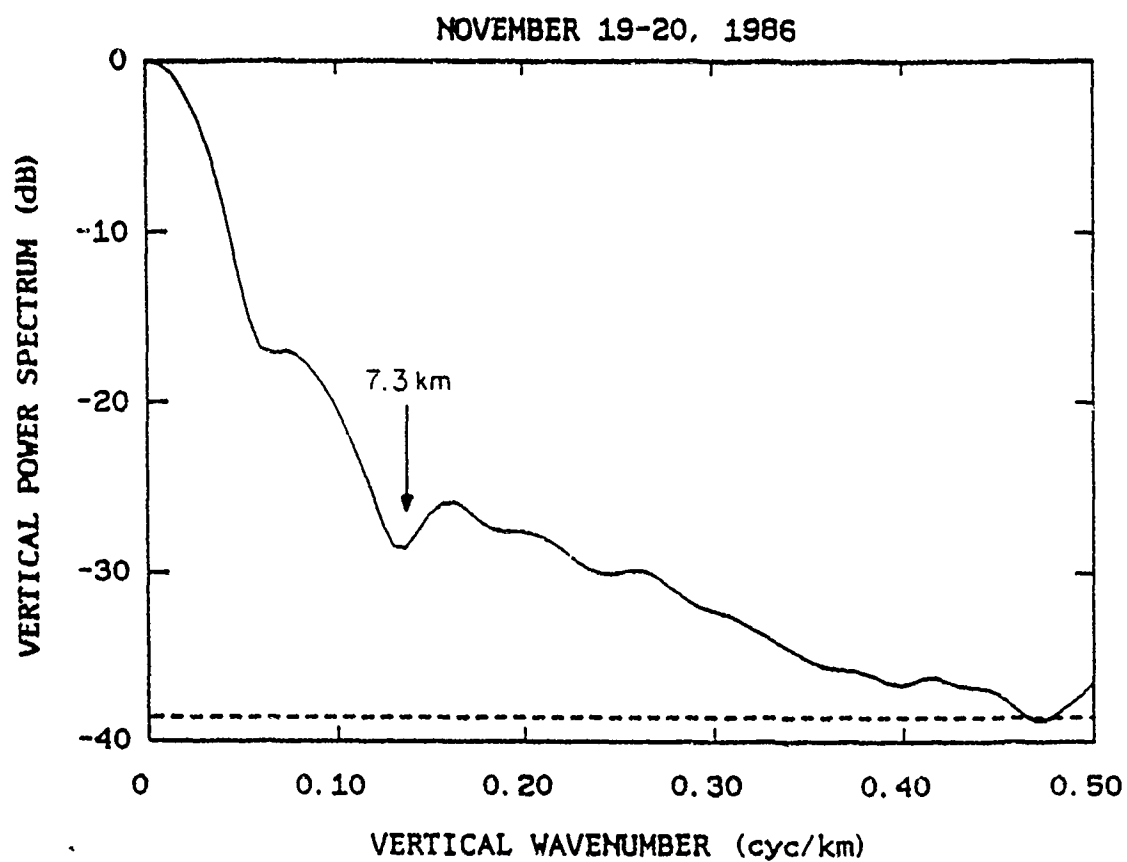


Figure 4.13. The average vertical wavenumber power spectrum computed for the Na density profiles collected from 0300 to 0625 MST on November 20, 1986.

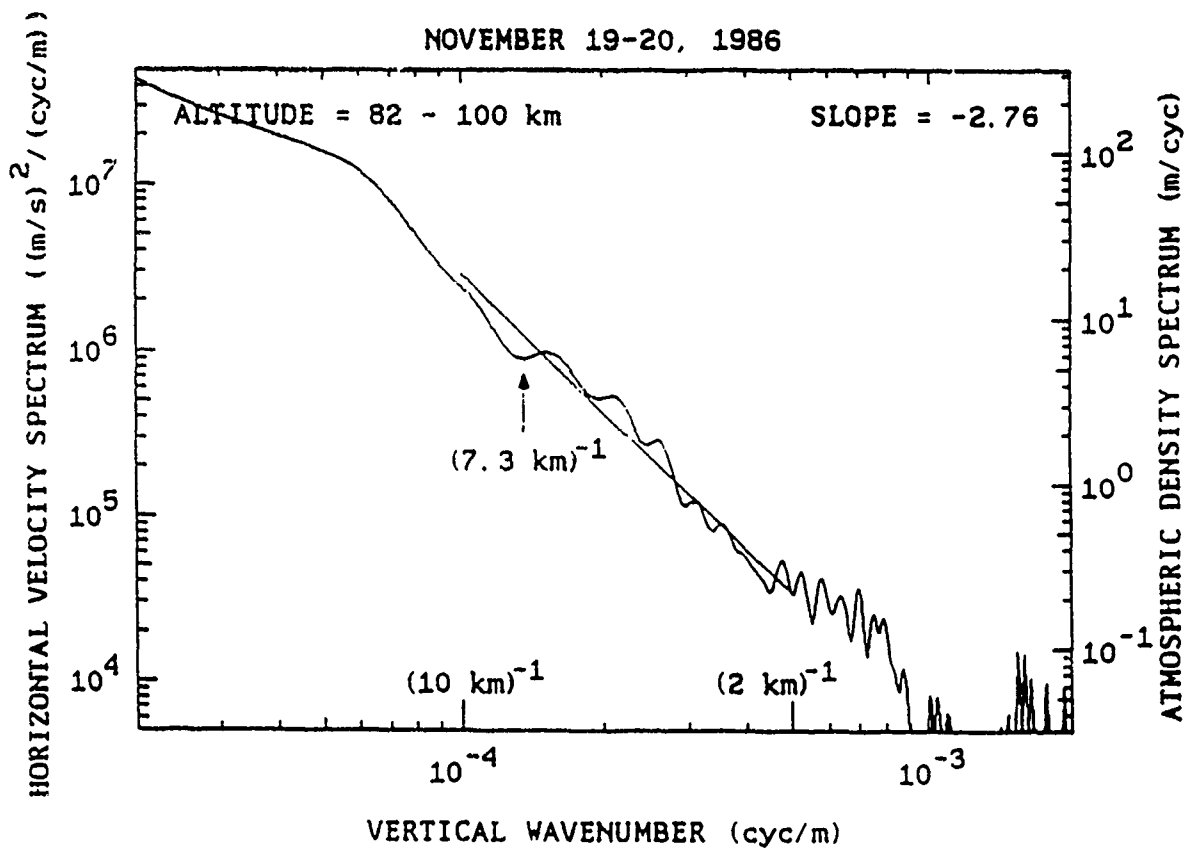


Figure 4.14. Vertical wavenumber spectrum measured on November 19-20, 1986. The Na density profiles were initially filtered temporally with a cutoff of 20 min. The straight line is a linear regression fit which was used to estimate the spectral slope over vertical scales ranging from 2 to 10 km.

level was estimated and subtracted from the spectrum. The results were plotted in Figure 4.14. The straight line is a linear regression fit which was used to estimate the spectral slope over the vertical wavelength range from 2 to 10 km. The slope is estimated to be about -2.76.

4.5 Airborne Lidar Observations

In this section, the data collected during the eastward flight of November 15-16, 1986 will be presented. As discussed earlier in Section 3.4, the Na layer was dominated by two quasi-monochromatic waves during the flight. These two waves appear to have induced the westward propagating centroid maxima and minima plotted in Figure 3.10. The longitudinal variations of the rms width and column abundance observed during the flight also exhibited the westward propagating maxima and minima as can be seen in Figure 4.15. The parameters of the two waves were summarized in Table 3.2. The intrinsic periods of these two waves were about 6 hours and 2 hours, which are surprisingly similar to those of the two dominant waves observed with the ground-based lidar at Broomfield, Colorado on November 19-20. The two waves observed during the flight were also dominant only on the bottomside of the layer. The zonal wavelength of the 6-hour period wave observed during the eastward flight was related to the horizontal separation distance between two centroid maxima measured on the eastbound leg near 99.5°W and 92.2°W in Figure 3.10, and also the distance between two maxima measured on the westbound leg near 104.6°W and 93.2°W. These centroid maxima were propagating westward at an apparent velocity of about 30 m s^{-1} . The centroid maximum near 104.6°W measured on the westbound leg would have propagated to Broomfield, Colorado (105°W) near 0300 MST on November 16. The centroid height observed at Broomfield on November 19-20 reached a maximum near 0500 MST (Figure 4.5), which may represent a time shift of 2 hours for the 6-hour period wave over the time span of 4 days from November 16 to 20 when the two observations were made.

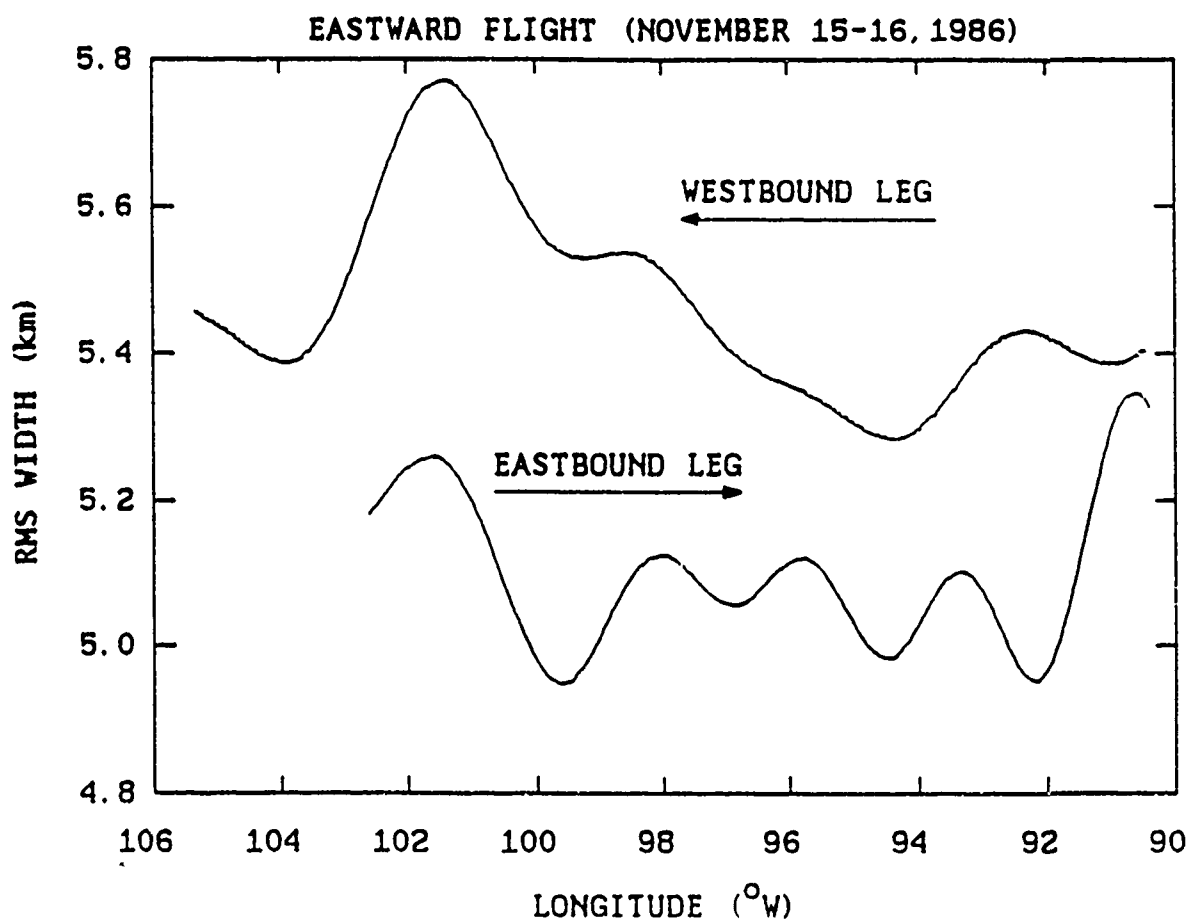


Figure 4.15 a). Longitudinal variations of the rms width of the Na layer observed during the eastward flight on November 15-16, 1986.

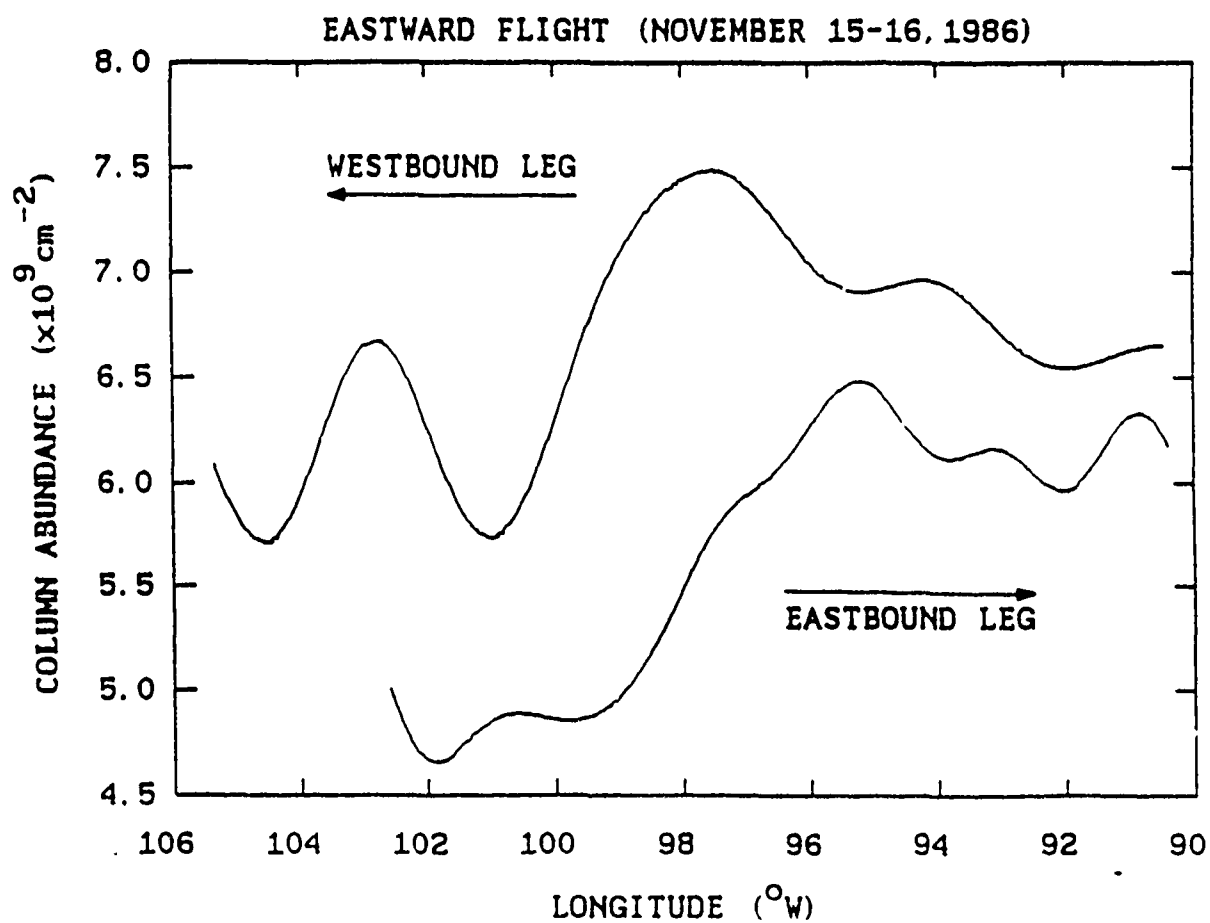


Figure 4.15 b). Longitudinal variations of the column abundance of the Na layer observed during the eastward flight on November 15-16, 1986.

The average vertical wavenumber power spectrum computed for the Na density profiles obtained during the westbound flight leg in the longitude range from 96°W to 102°W is plotted in Figure 4.16. The vertical wavelength of a dominant wave in the spectrum is 7.3 km, and is believed to be the vertical wavelength of the 6-hour period wave. This wavelength is the same as that estimated for the 6-hour period wave observed with the ground-based lidar on November 19-20.

The vertical wavelength of the 2-hour period wave observed with the ground-based lidar was estimated to be 12 km. The vertical wavelength of the 2-hour period wave observed with the airborne lidar is related to the intrinsic horizontal wavelength by the gravity wave dispersion relation,

$$\lambda_z = \frac{T_B}{T_{in}} \lambda_x \quad (4.4)$$

where λ_z = intrinsic vertical wavelength, and

λ_x = intrinsic horizontal wavelength.

The intrinsic horizontal wavelength of the 2-hour period wave was not measured during the flight. However, by using the intrinsic zonal wavelength of 263 km and the intrinsic period of 102 min, the vertical wavelength is calculated to be 12.9 km, which is quite comparable to the vertical wavelength of 12 km for the 2-hour period wave observed with the ground-based lidar on November 19-20. This appears to indicate that the intrinsic zonal wavelength measured during the flight is comparable to the intrinsic horizontal wavelength, and consequently the wave was propagating almost zonally (westward).

4.6 Summary

The results of the joint lidar/radar campaign conducted in November, 1986 are presented. The lidar observations were obtained with the airborne Na lidar and with the ground-based lidar at Broomfield and Denver, Colorado. The airborne lidar data presented in this chapter were collected during the eastward flight on the night of November 15, 1986, and the ground-based

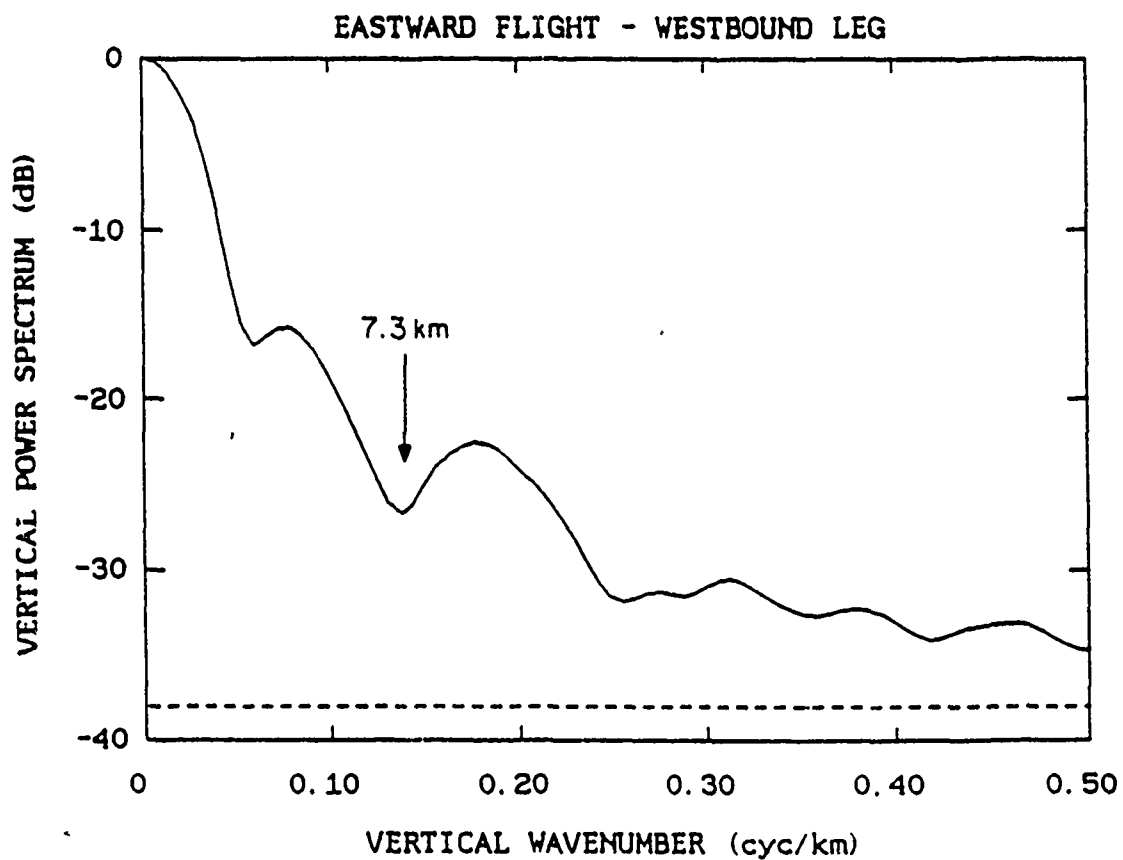


Figure 4.16. The average vertical wavenumber power spectrum computed for the Na density profiles collected during the westbound leg of the eastward flight in the longitude range from 96°W to 102°W.

lidar data were obtained for a total of 8 hours on the night of November 19, 1986. The radar observations were obtained with the ST radar at Platteville, Colorado during the period from November 4 to 20, 1986.

The ground-based lidar observations showed very strong density perturbations in the Na layer with a period of approximately 6 hours. The airborne observations of the Na layer also exhibited a dominant variation by a wave with a period of approximately 6 hours. The intrinsic zonal wavelength of this wave was 772 km, and the intrinsic zonal phase velocity was 35 m s^{-1} . This wave appears to be responsible for the 6-hour period density variations observed with the ground-based lidar. The 6-hour period variations observed with both the ground-based and airborne lidars were dominant only on the bottomside of the layer (80-90 km). The radar observations also showed a dominant horizontal wind variation with a period of 6 hours. In the altitude region which corresponds to the bottomside of the Na layer, the amplitudes of the 6-hour period zonal winds were much larger than the amplitudes of the 6-hour period meridional winds. The 6-hour period zonal winds were most dominant at the altitude of 88 km, and the amplitude at this altitude was approximately 9 m s^{-1} . The horizontal wind amplitude of the 6-hour period wave observed with the ground-based lidar was estimated from the vertical wavenumber power spectrum to be 17 m s^{-1} .

Both the ground-based and airborne lidar observations revealed another strong variation in the Na layer with a period of about 2 hours. The intrinsic zonal wavelength of the 2-hour period wave observed during the eastward flight was 263 km, and the intrinsic zonal phase velocity was 43 m s^{-1} . The vertical phase velocity of the 2-hour period wave observed with the ground-based lidar was 1.8 m s^{-1} , and the vertical wavelength was estimated to be 12 km. These 2-hour period waves were again dominant only on the bottomside of the Na layer, and believed to be propagating westward.

5. LIDAR OBSERVATIONS OF SPORADIC SODIUM LAYERS AT MAUNA KEA OBSERVATORY , HAWAII

5.1 Introduction

The upper atmospheric Na layer has been explored with lidar systems for almost two decades. The layer is generally confined to an altitude range of 80-110 km, with a peak density near 92 km of about 10^3 - 10^4 cm⁻³. Meteoric ablation is believed to be the dominant source of the Na layer [*Jegou et al.*, 1985a]. Lidar studies have shown that the layer is an excellent tracer of wave motions. The layer profile is particularly sensitive to gravity and tidal wave effects, because the wave amplitudes are usually large near the mesopause and the steep Na density gradients on the bottom and topsides of the layer tend to enhance the observed wave perturbations. An extensive analysis of the monochromatic gravity waves observed in the Na layer above Urbana, Illinois (40°N, 88°W), was recently reported by *Gardner and Voelz* [1987]. A study of tidal waves observed above the same location has also been reported by *Kwon et al.* [1987]. In addition to exploring wave dynamics, lidar studies have also shown that the layer can be used to measure the temperature of the upper mesosphere and lower thermosphere. The temperature profiles are obtained from measurements of the Doppler-broadened resonance line width of the Na atoms [*Gibson et al.*, 1979]. Recently, *Neuber et al.* [1988] reported extensive wintertime measurements of the temperature profile near the mesopause, using a Na lidar at Andoya, Norway (69°N, 16°E).

Lidar measurements at low- and high-latitude sites have occasionally exhibited the sporadic developments of very dense narrow Na layers. In the late 1970s, *Clemesha et al.* [1978] were the first to report such observations at Sao Paulo, Brazil (23°S, 46°W). More recently, the observations of *von Zahn et al.* [1987] and *von Zahn and Hansen* [1988] at Andoya, Norway, have generated considerable interest in this intriguing phenomenon. *Gardner et al.* [1988] have also recently reported similar measurements of sporadic Na layers above

Longyearbyen, Svalbard (78°N, 15°E). The peak densities of these sporadic layers sometimes exceed 10^4 cm^{-3} and can be as much as 10 times higher than the peak densities of the normal Na layer. The thicknesses of the sporadic layers are very small, typically of the order of 1 km FWHM. Because development times are usually less than a few tens of minutes, these layers have also been referred to as "sudden layers" [von Zahn *et al.*, 1987]. In this chapter, the characteristics of 16 sporadic Na layers observed at the low-latitude site of Mauna Kea, Hawaii (19°50'N, 155°28'W), during 5 nights of observations in January 1987 will be discussed.

5.2 Observations

From January 17 to 22, 1987, the University of Illinois (UIUC) lidar system was used to study the Na layer above Mauna Kea Observatory. The temporal resolution of the lidar was 100 s, and the vertical resolution was 150 m. All the measurements were made at zenith, and the laser beam divergence was 0.4 mrad full width at the e^{-2} intensity point. The beam illuminated a horizontal circular area in the Na layer, with a diameter of about 40 m. The lidar system is described in more detail by *Thompson and Gardner* [1987]. During the 5 nights of the campaign, 30 hours of Na measurements were obtained, and a total of 16 sporadic layers were observed. The system operation periods and statistics of the sporadic layers are listed in Table 5.1. At least one sporadic layer developed during each night, except on January 17, when only 1.5 hours of observations were obtained. On the night of January 21, nine sporadic layers developed within a period of 2 hours, between 2100 and 2300 LST. To illustrate the behavior of these sporadic layers, a sequence of density profiles is plotted in Figure 5.1. A normal density profile, obtained at 2133 LST, is plotted in Figure 5.1a. A small sporadic Na enhancement first appeared near 105 km at 2135 LST. Then, four narrow layers formed quickly between 2200 and 2217 LST in the altitude region between 95 and 100 km. The density profile obtained at 2218 LST, including two of these sporadic layers, is plotted in Figure 5.1b. The layer at 97 km moved downward with an apparent velocity of 2.7 m s^{-1} and merged with the lower layer near

**Table 5.1. Statistics of Na Sporadic Layers Observed
at Mauna Kea, Hawaii**

Date	Measurement Time (LST)	Measurement Duration (hour:min)	Number Sporadic Layers
January 17, 1987	0127 - 0252	1:25	0
January 18, 1987	0153 - 0620	4:27	1
January 19, 1987	2115 - 2118	0:03	1
January 20, 1987	0053 - 0622	5:29	2
January 20-21, 1987	2052 - 0618	9:26	2
January 21-22, 1987	2059 - 0601	9:02	10
	Total	29:53	16

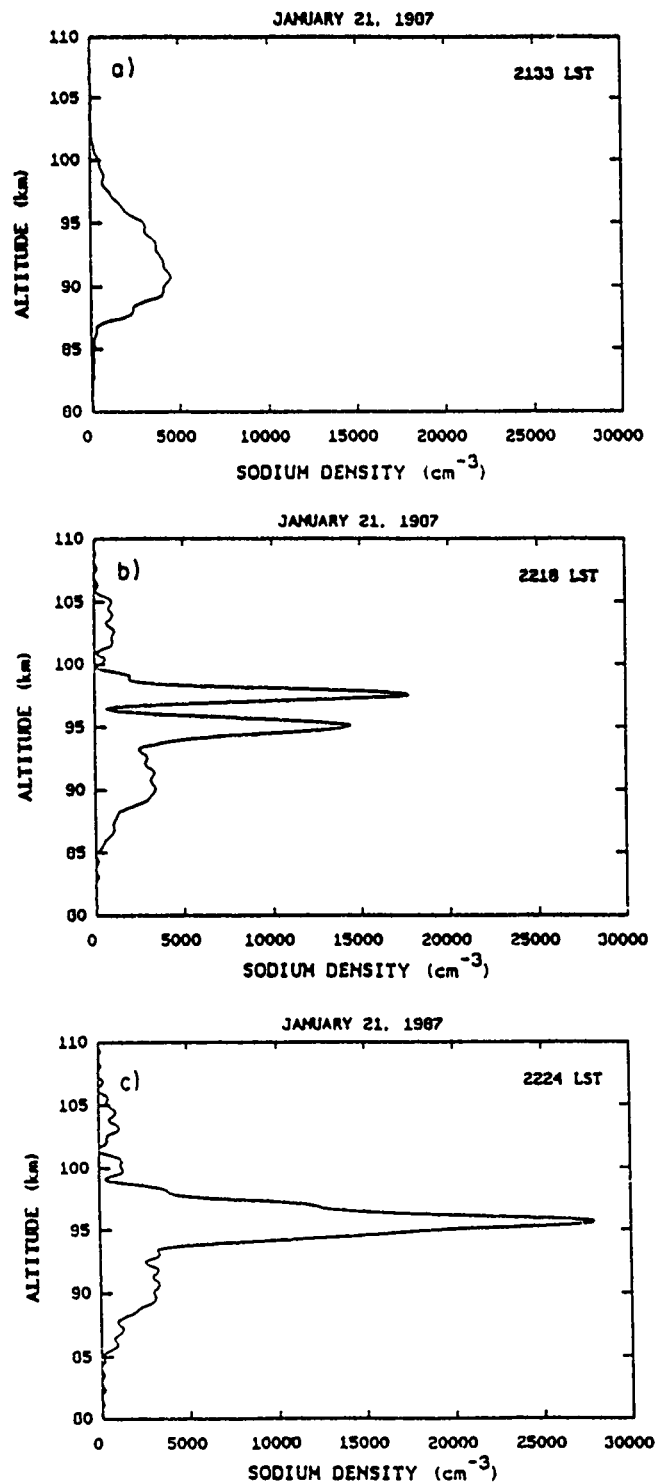


Figure 5.1. Sodium density profiles measured at a) 2133 LST, b) 2218 LST, and c) 2224 LST at Mauna Kea Observatory on January 21, 1987.

95 km at 2224 LST, as seen in Figure 5.1c. The lower layer reached a maximum density of $2.8 \times 10^4 \text{ cm}^{-3}$, with a thickness of 2.0 km FWHM. Although the peak density decreased after 2224 LST, this layer was observed continuously for almost 8 hours. During the night, additional sporadic layers formed near 103 km, as seen in Figure 5.2. The dominant sporadic layer moved downward from about 96 to 87 km, with a remarkably constant velocity of 31.4 cm s^{-1} . The altitude of this layer is plotted versus time in Figure 5.3. The straight line drawn in Figure 5.3 is a linear regression fit, which was used to estimate the average vertical velocity. When the layer descended to an altitude of 91 km at 0255 LST, the peak density began increasing from approximately 3000 cm^{-3} , reaching a maximum density of about 9280 cm^{-3} at 0407 LST, as seen in Figure 5.4.

The peak density of the dominant layer is plotted versus time in Figure 5.5. When the peak density first reached the maximum at 2224 LST, the column abundance of the sporadic layer was $6.7 \times 10^9 \text{ cm}^{-2}$, a value which was 74% of the column abundance of the entire Na layer. About 13 min before the sporadic layer began forming, the Na abundance in the region of the layer was only $1.1 \times 10^9 \text{ cm}^{-2}$, and the column abundance of the entire Na layer was $3.5 \times 10^9 \text{ cm}^{-2}$. Thus when it was most prominent, the sporadic layer contained a substantial amount of Na, which almost doubled the abundance of the entire layer. The rapid growth and decay of the first maximum of this sporadic layer lasted for only 40 min. The average Na production and decay rates were 28.4 and $12.6 \text{ cm}^{-3} \text{ s}^{-1}$, respectively. The maximum production and decay rates were 60.9 and $68.0 \text{ cm}^{-3} \text{ s}^{-1}$, respectively. Recently, *von Zahn and Hansen* [1988] defined a parameter, called the strength factor, to classify the sporadic layers that they observed. The strength factor is defined as the ratio of the maximum peak density of the sporadic layer to the density of the normal layer at the altitude of the peak of the sporadic layer. The strength factor of the sporadic layer illustrated in Figure 5.1c was 13.8. The second rapid enhancement of this layer occurred about 6 hours after the initial formation. The average Na production and decay

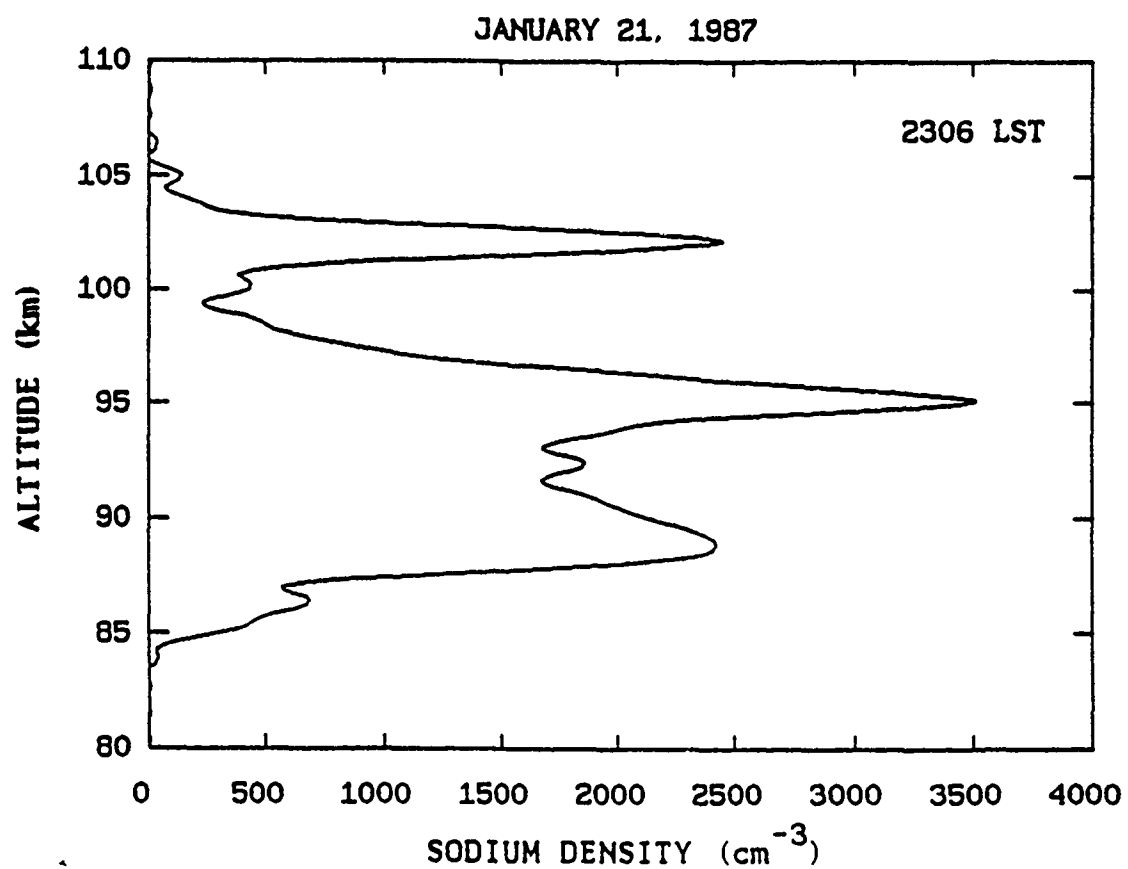


Figure 5.2. Sodium density profile measured at 2306 LST on January 21, 1987 at Mauna Kea Observatory.

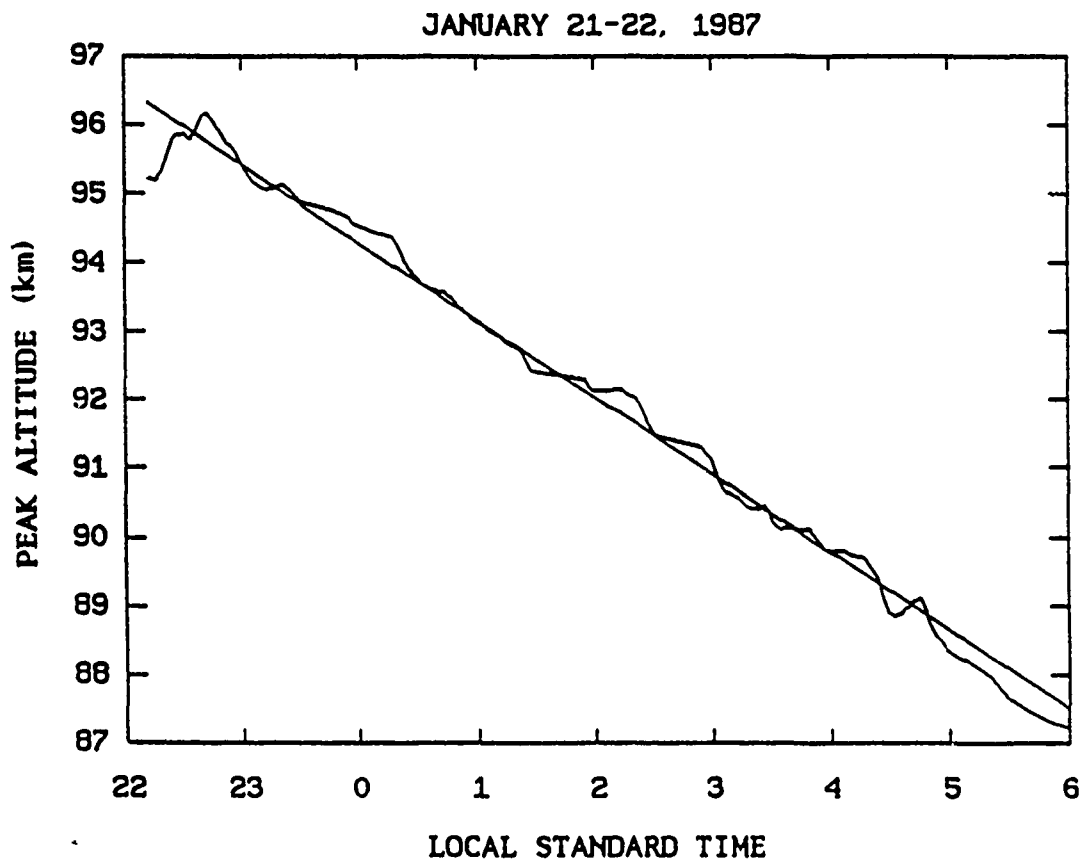


Figure 5.3. Temporal variation of the altitude of the dominant sporadic Na layer observed on January 21-22, 1987 at Mauna Kea Observatory. The straight line is a linear regression fit which was used to estimate the average vertical velocity.

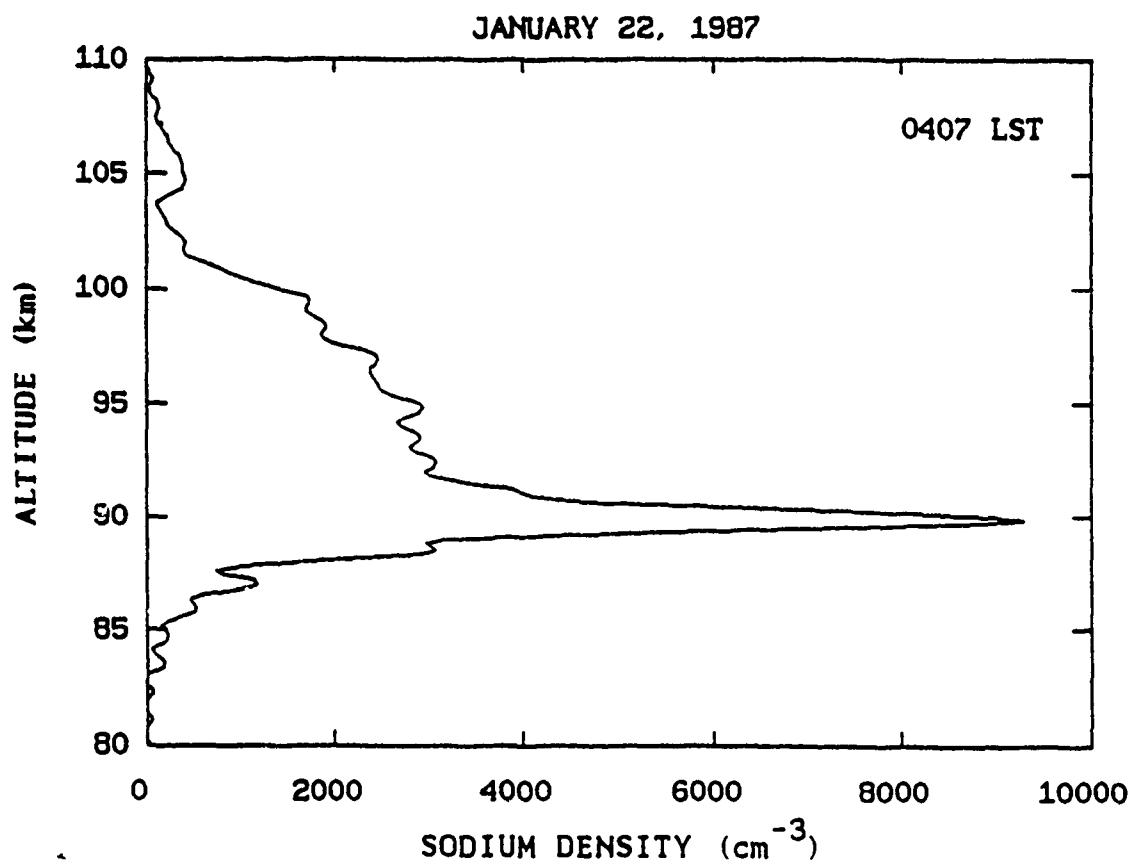


Figure 5.4. Sodium density profile measured at 0407 LST on January 22, 1987 at Mauna Kea Observatory.

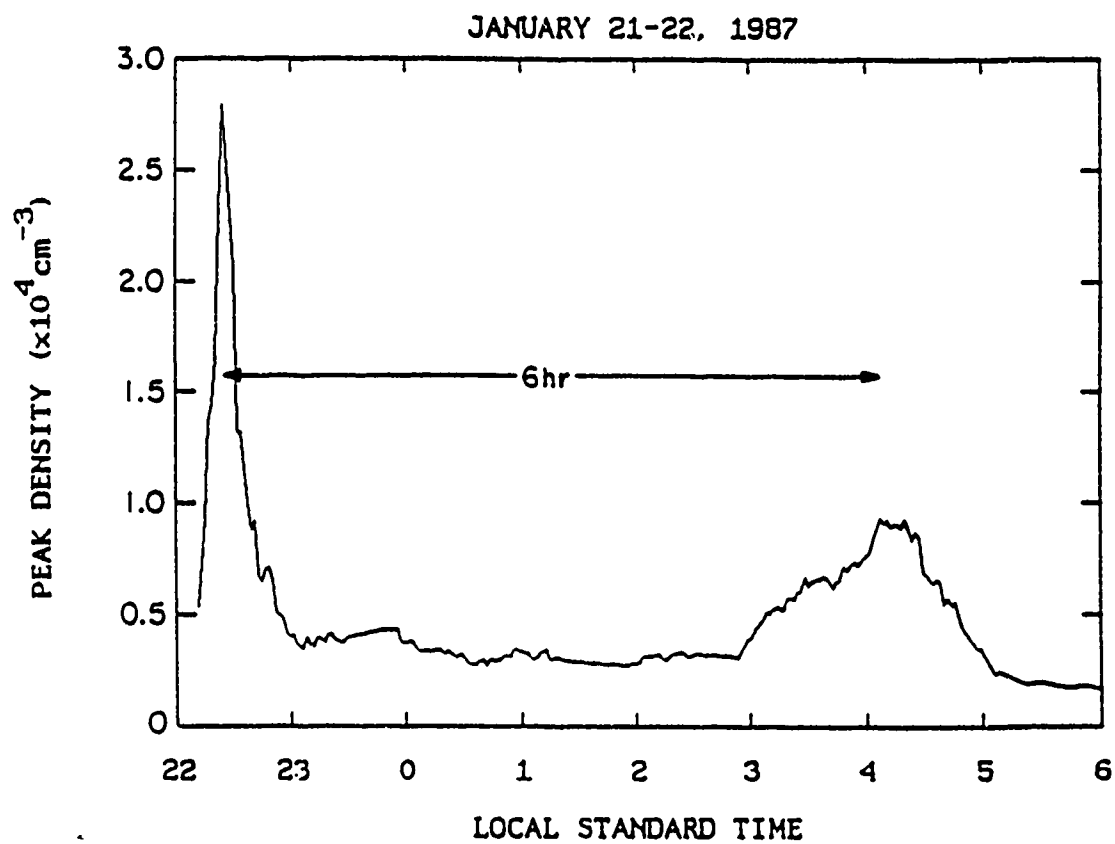


Figure 5.5. Temporal variation of the density at the peak of the dominant sporadic Na layer observed on January 21-22, 1987 at Mauna Kea Observatory.

rates for the second maximum were 1.0 and $2.2 \text{ cm}^{-3} \text{ s}^{-1}$, respectively, and the strength factor was 6.8 .

The full, top-half, and bottom-half widths of the dominant sporadic layer are plotted versus time in Figure 5.6. The widths were determined at 80% of the peak density. When examining these widths, it is helpful to remember that the full width determined at the 80% points is 1.34 times larger than the rms width of a Gaussian layer. The full width was very narrow, of the order of 1 km, during the development periods of the two maximum peak densities. Then, as the peak density decreased, the full width became larger. The average full width measured over the total observation period was 1.25 km. The increase of the full width from 0100 to 0300 LST was largely due to an increase of the top-half width. During this period the full width increased at the rate of 14.6 cm s^{-1} , the top-half width increased at the rate of 11.9 cm s^{-1} , and the bottom-half width increased at only 2.7 cm s^{-1} . After the second maximum peak density near 0410 LST, both top- and bottom-half widths began growing, resulting in a relatively fast growth of the full width. The average top- and bottom-half widths over the total observation period were 0.71 and 0.54 km, respectively.

Although the broadening of the sporadic layer might suggest that the Na diffused into the normal layer, most of the Na atoms apparently disappeared as the peak density of the sporadic layer decreased. The column abundance of the entire Na layer observed on the night of January 21 is plotted in Figure 5.7. The maximum near 2220 LST occurred when the peak density of the sporadic layer was at its first maximum. After this maximum the column abundance decreased rapidly as the peak density of the sporadic layer decreased. Therefore it appears that most of the Na in the sporadic layer simply disappeared and was not redistributed throughout the rest of the Na layer.

In addition to the dominant layer, another sporadic layer formed near 103 km at 2251 LST (Figure 5.2). The peak density and altitude of this sporadic layer are plotted versus time in Figure 5.8. The highest peak density was 3390 cm^{-3} . The strength factor was 16.3, a value

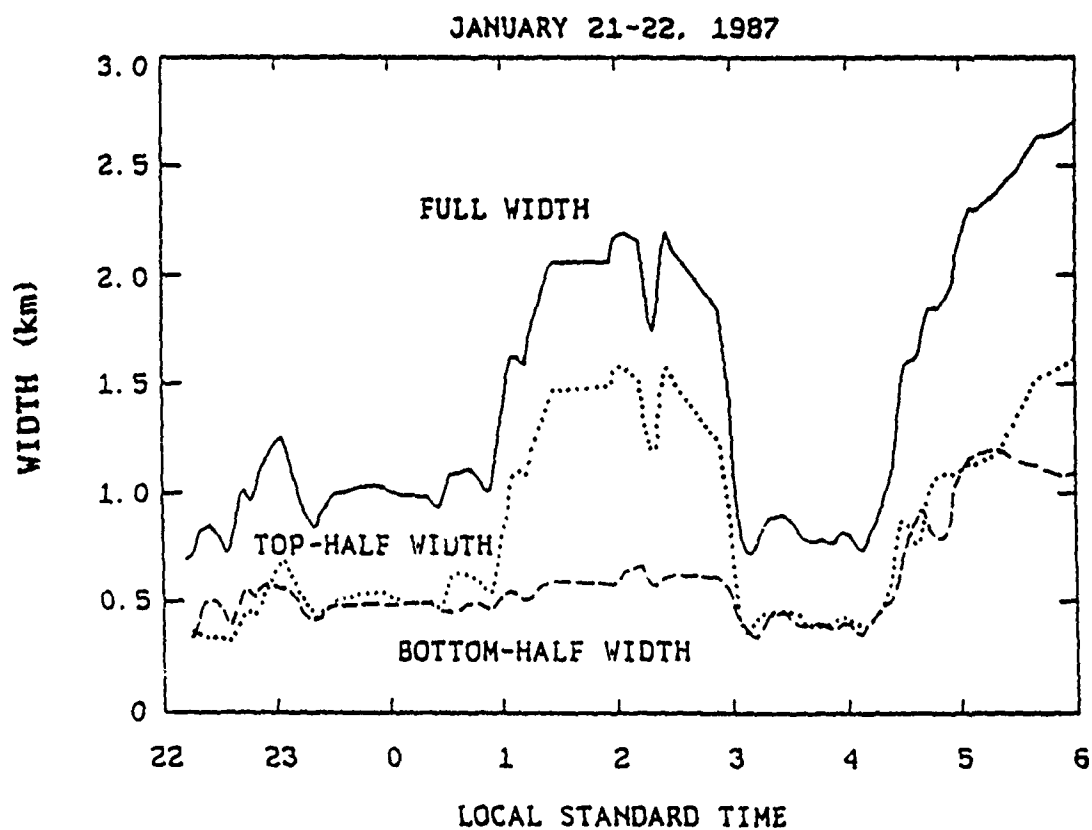


Figure 5.6. Temporal variations of the widths of the dominant sporadic Na layer observed on January 21-22, 1987 at Mauna Kea Observatory.

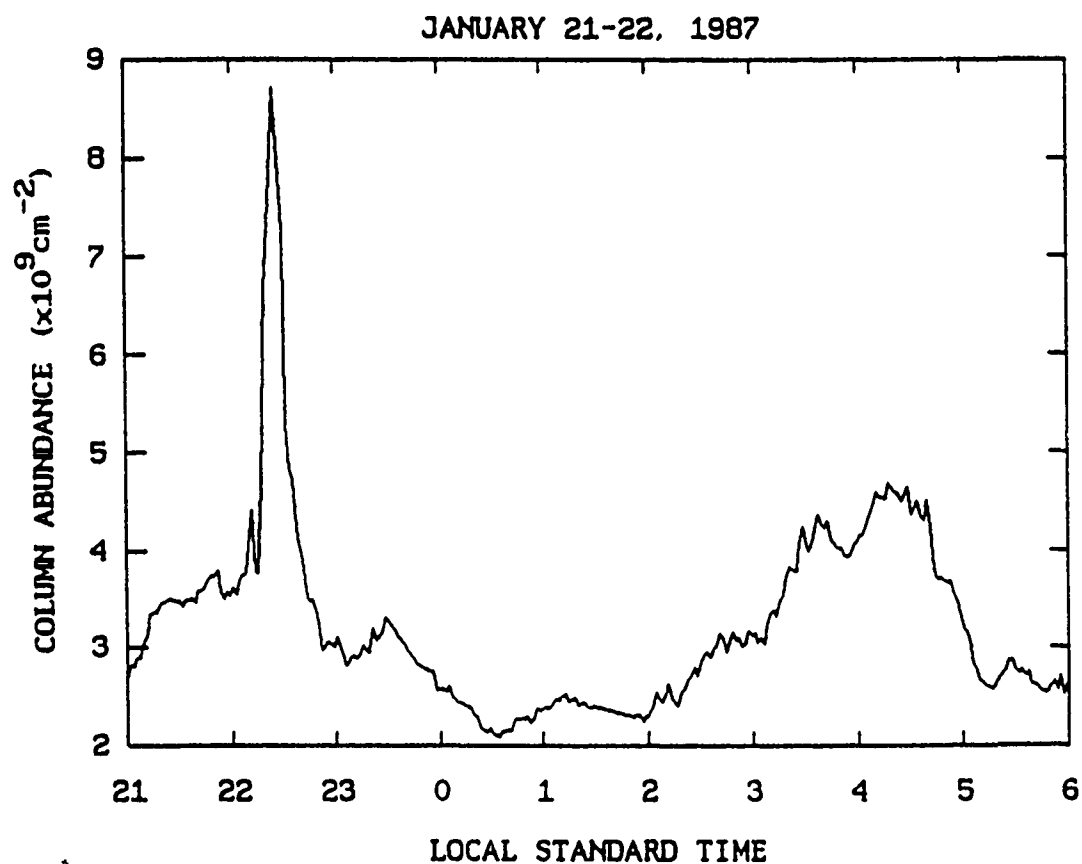


Figure 5.7. Temporal variation of the column abundance measured on January 21-22, 1987 at Mauna Kea Observatory.

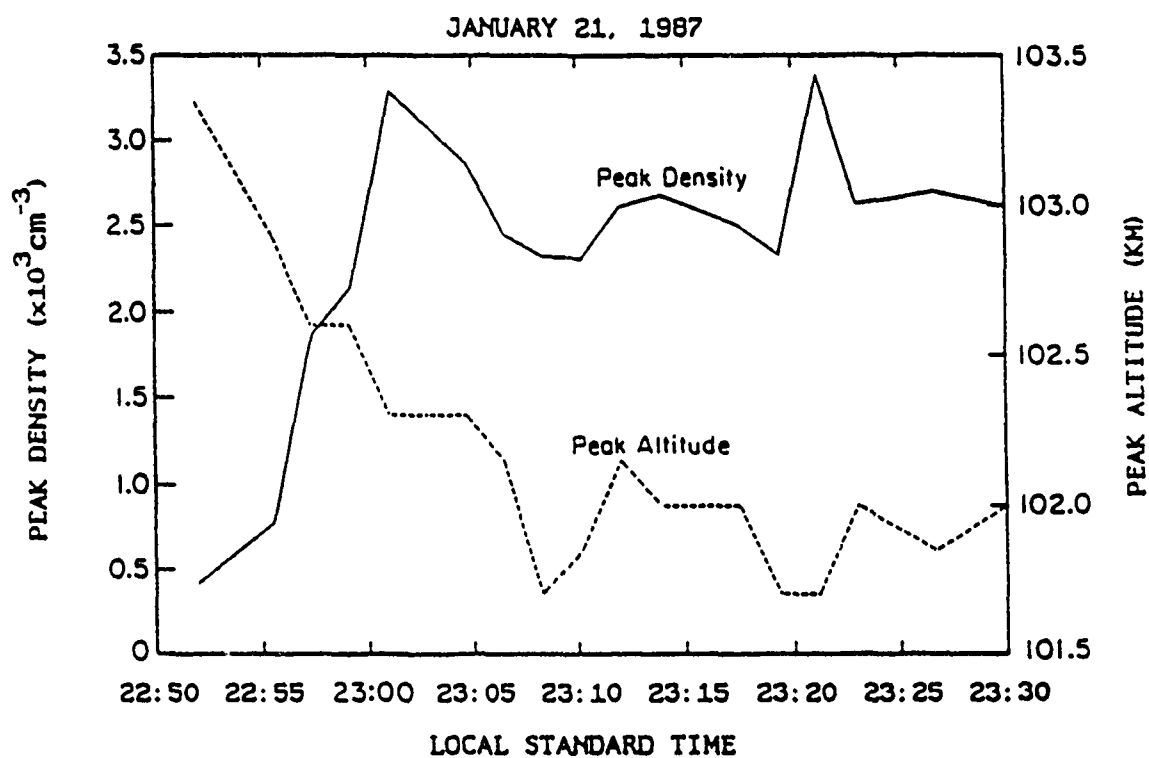


Figure 5.8. Temporal variations of the peak density and altitude of the high altitude sporadic Na layer observed on January 21, 1987 at Mauna Kea Observatory.

which is higher than that of the dominant sporadic layer. The higher strength factor was due to the very small density of the normal Na layer at this high altitude. From 2251 to 2301 LST the peak density of this sporadic layer increased rapidly as the peak altitude moved downward. Then, from 2301 to 2330 LST, the peak density stabilized at a value near 2800 cm^{-3} as the peak altitude also stabilized near 102 km. During the rapidly changing period the Na production rate and vertical velocity were $5.2 \text{ cm}^{-3} \text{ s}^{-1}$ and -1.8 m s^{-1} , respectively. During the stable period the Na density decreased slightly. The decay rate and vertical velocity of the layer were $0.1 \text{ cm}^{-3} \text{ s}^{-1}$, and -22.8 cm s^{-1} , respectively.

Another sporadic layer with a high peak density formed near 97 km at 2217 LST (Figure 5.1b). This layer lasted for 6 min and was present in only four density profiles. The layer moved downward rapidly and merged with the dominant layer at 95 km. The apparent vertical velocity of the peak was remarkably constant at -2.7 m s^{-1} . The altitude of the layer is plotted versus time in Figure 5.9. The four measurement points are represented by the crosses. The highest peak density of this short-lived layer was $17,680 \text{ cm}^{-3}$, and the maximum Na production rate was $92.7 \text{ cm}^{-3} \text{ s}^{-1}$.

The 16 sporadic layers are organized in decreasing order of the highest peak densities and then numbered. The characteristics of each layer are listed in Appendix 3. The characteristics of the dominant sporadic layer of January 21-22 are divided into two groups (1 and 5), corresponding to the two development periods near 2200 and 0300 LST.

The most significant pattern of the sporadic layers appears to be the occurrence times. Lidar observations were conducted near 0300 LST on 4 days of the campaign. During 3 of these 4 days, four sporadic layers (5, 7, 11, and 14) began forming within the short time span of 15 min from 0253 to 0308 LST. Even on January 21, when narrow sporadic layers did not form during this period, a rapid increase of Na density between 90 and 95 km was observed near 0300 LST. The altitudes of the sporadic layers are plotted versus time in Figure 5.10. A total of 10 sporadic layers began forming between 2100 and 2300 LST, but no layers formed between

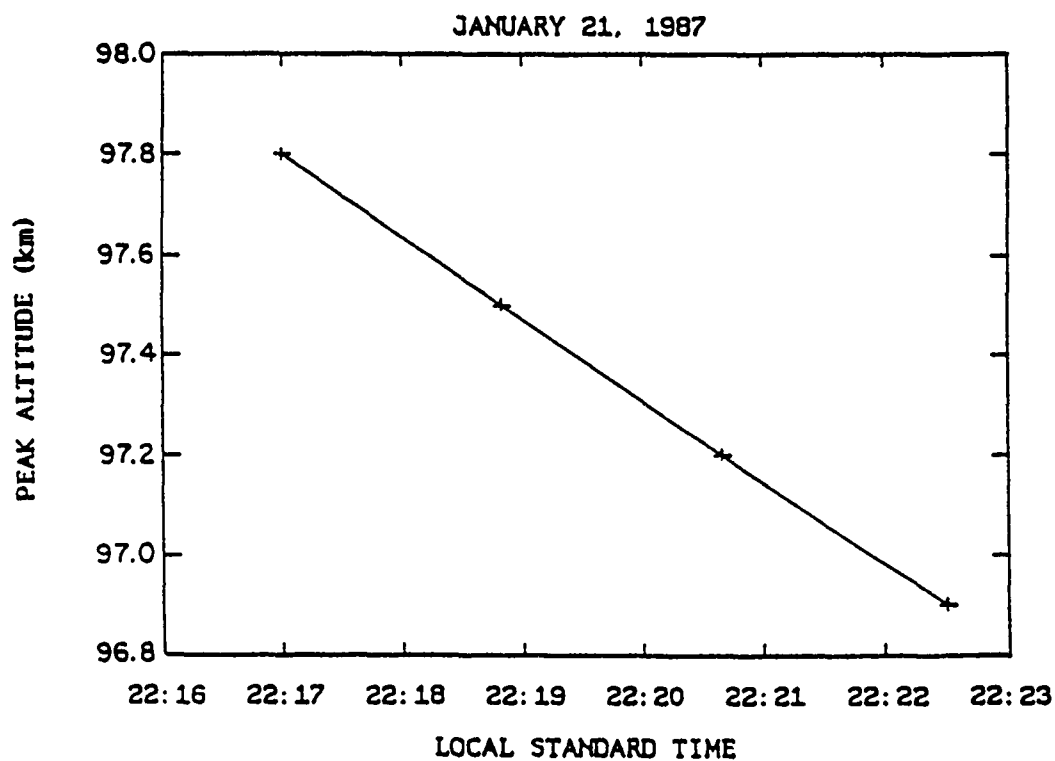


Figure 5.9. Temporal variation of the altitude of the short-lived sporadic Na layer observed on January 21, 1987 at Mauna Kea Observatory. Crosses represent the four measurement points.

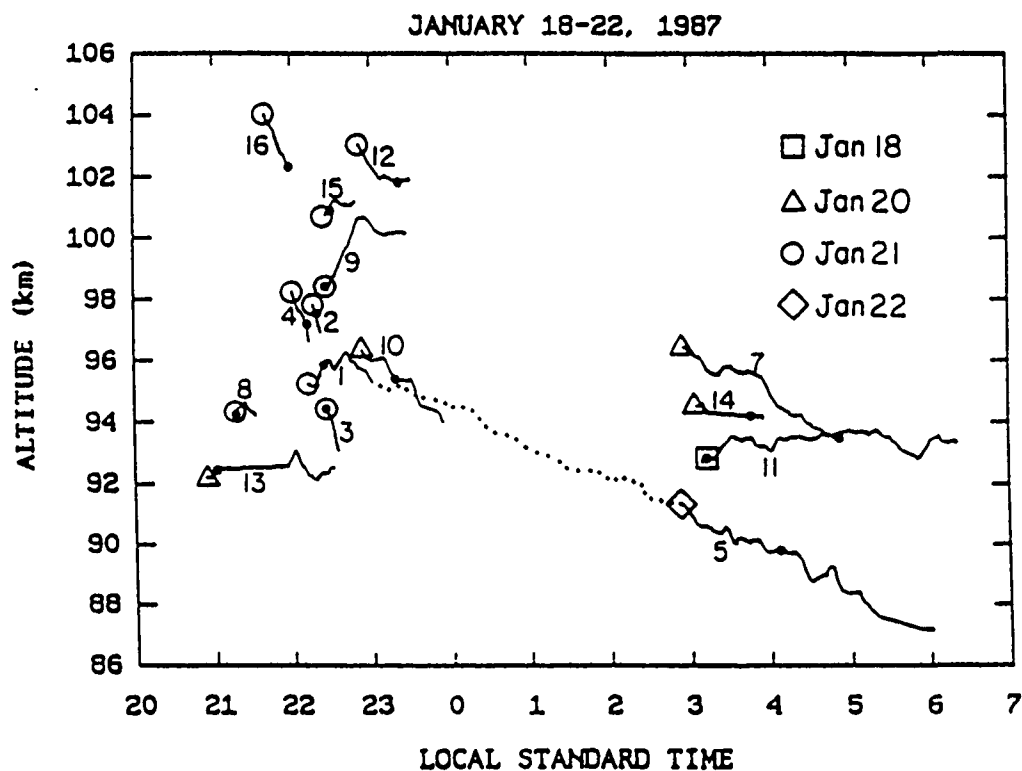


Figure 5.10. Temporal variations of the altitudes of the sporadic Na layers measured from January 18 to 22, 1987 at Mauna Kea Observatory. Dots represent the times and altitudes of the maximum peak densities. Squares, triangles, circles, and diamonds represent starting times and altitudes. The sporadic layers are numbered in accordance with Appendix 3.

2300 and 0250 LST. In Table 5.2 the characteristics of the layers which formed between 2100 and 2300 LST are compared with those of the layers which formed near 0300 LST. The difference in the mean times of the maximum peak densities of the two groups of the layers was about 6 hours. In general, the layers of the early morning had lower starting altitudes, longer durations, and longer formation and dissipation periods than the layers of the late evening. A total of nine sporadic layers began forming in the altitude range between 94 and 99 km, two between 91 and 93 km, and three between 100 and 104 km. The highest and lowest peak altitudes were 104.4 km (layer 16) and 86.6 km (layer 5), respectively. A total of 10 layers moved downward, four moved upward, and one moved upward and then downward. Three longest duration layers (1, 5, and 7) moved downward, with mean apparent velocities of about 30 cm s^{-1} . Six layers moved at very high velocities, ranging from 1.6 to 2.9 m s^{-1} . All six of these layers were observed on January 21 when the dominant layer developed. In general, the top-half widths were larger than the bottom-half widths, and the broadening speeds of the top-half widths were also considerably faster than those of the bottom-half widths. The formation periods were generally shorter than the decay periods.

The general characteristics of the prominent sporadic layers observed in Sao Paulo, Brazil, Andoya, Norway, and Hawaii are compared in Table 5.3. The prominent layer of Sao Paulo is chosen from two reported events in the literature [Clemesha *et al.*, 1978; 1980], and that of Andoya from 10 reported events [von Zahn and Hansen, 1988]. The prominent sporadic layers above Hawaii and Sao Paulo were observed during winter periods, whereas those above Andoya were observed during summer. Most of the characteristics are surprisingly similar. In fact, the starting altitudes are $\sim 95 \text{ km}$, and peak densities and vertical velocities are all comparable.

Table 5.2. Comparison of Sporadic Na Layers Observed in the Late Evening and Early Morning at Mauna Kea, Hawaii

	Late Evening Occurrences	Early Morning Occurrences
Number of Observations	12	4
Mean Starting Time	2207 LST (± 38 min)	0301 LST (± 7 min)
Mean Starting Altitude (km)	97.8 (± 3.8)	93.8 (± 2.2)
Mean Time of Maximum Density	2217 LST (± 42 min)	0359 LST (± 42 min)
Mean Altitude of Maximum Density (km)	97.2 (± 3.2)	92.6 (± 1.9)
Mean Duration (min) ^a	55.2 (± 73.3)	137.5 (± 72.5)
Mean Formation Period (min)	6.1 (± 4.1)	33.9 (± 30.3)
Mean Dissipation Period (min)	18.3 (± 18.3)	51.8 (± 30.8)

^aIn some cases duration was larger than sum of formation and dissipation periods because maximum density of the sporadic layer was maintained for several minutes.

Table 5.3. Characteristics of the Most Prominent Sporadic Na Layers Reported in Literature

Location	Sao Paulo, Brazil ^a (23°S, 46°W)	Andoya, Norway ^b (69°N, 16°E)	Mauna Kea, Hawaii ^c (20°N, 155°W)
Date	Aug. 26, 1979	Aug. 20, 1986	Jan 21-22, 1987
Starting Time	0100 LST	2209 LST	2211 LST
Absolute Peak Density	$4 \times 10^4 \text{ cm}^{-3}$	$2.3 \times 10^4 \text{ cm}^{-3}$	$2.8 \times 10^4 \text{ cm}^{-3}$
Strength Factor	12	24	14
Width	2 km	0.8 km FWHM	2 km FWHM
Formation Period	15 min	7 min	13 min
Duration of Sporadic Layer	30 min	4 hour	8 hour
Starting Altitude	95 km	95 km	95 km
Vertical Velocity of Peak	NA	-28 cm s ⁻¹	-31 cm s ⁻¹

NA, not available.

^aFrom *Clemesha et al.* [1980].

^bFrom *von Zahn et al.* [1987] and *von Zahn and Hansen* [1988].

^cFrom this paper.

5.3 Discussion

The mechanisms responsible for creating the sporadic Na layers have been the subject of much speculation. On the one hand, *von Zahn and Hansen* [1988] suggested that the sporadic layers were formed when auroral excitation caused Na to be evaporated from the surfaces of mesospheric dust particles. On the other hand, *Clemesha et al.* [1980] suggested that the sporadic layers were vapor clouds created by meteoric ablation. Both groups noted that the prevailing winds would rapidly transport the sporadic layers horizontally. By making measurements with a steerable lidar system and two photometers located at two sites separated by 107 km, *Clemesha et al.* [1980] estimated that the horizontal velocities of the sporadic layers observed at Sao Paulo were of the order of 200 m s^{-1} . The photometer measurements exhibited increases in Na D-line intensities during the occurrence of the sporadic layers. However, no changes in atomic oxygen and hydroxyl emissions were observed. In November 1987, increases of Na D-line intensities were also measured at Longyearbyen, Svalbard, during the occurrence of a sporadic Na layer that was observed with the UTUC lidar [*Gardner et al.*, 1987, and Roger Smith, University of Alaska, private communication, 1987].

The sporadic Na layers appear to be related to sporadic E layers. Simultaneous occurrences of sporadic E and Na layers at almost identical altitudes were reported by both *von Zahn and Hansen* [1988] and *Clemesha et al.* [1980]. Both groups argued that the mechanisms creating the sporadic E layers were also responsible for the sporadic Na layers. However, the argument does not explain completely the sporadic E and Na layers observed in Hawaii. During the Na lidar campaign, an ionosonde was operated on Maui ($28^{\circ}48'N$, $156^{\circ}30'W$), which is located about 150 km northwest of the lidar site. Ionograms were routinely recorded every 15 min. On a total of eight occasions, sporadic E layers formed above Maui at almost the same times and altitudes as the sporadic Na layers above the lidar site at Mauna Kea. Seven of these eight sporadic Na layers formed within a short period of 50 min on the night of January 21, in the altitude region from 94.5 to 103.35 km. However, during the occurrences of the remaining

eight sporadic Na layers, no sporadic E layers were observed. For example, sporadic E layers appeared during the development period of the dominant sporadic Na layer of January 21, but no sporadic E layers appeared when the density of this Na layer began increasing to its second maximum near 0300 LST. In fact, for the four sporadic Na layers that formed in the early morning near 0300 LST no corresponding sporadic E layers were observed. Furthermore, sporadic E layers often formed above Maui when no sporadic Na layers were observed above Mauna Kea.

The winds induced by diurnal tides appear related to the formation of some of the sporadic Na layers observed in Hawaii. The three longest duration sporadic Na layers (1, 5, and 7 in Appendix 3) had average downward vertical velocities of 28, 33, and 35 cm s⁻¹. During winter the diurnal tide has been observed to be dominant at the low-latitude sites of Punta Borinquen (18°N) [Bernard *et al.*, 1981] and Arecibo (18°N) [Mathews, 1976]. The estimated vertical wavelengths of the diurnal tides range from 25 to 30 km at the altitudes of the Na layer. The corresponding vertical phase velocities of the diurnal tide are estimated to range between 29 and 35 cm s⁻¹. The three long-lived sporadic layers observed at Mauna Kea appear to move downward with these same velocities. Another indication that there is a link between the tidally induced winds and the sporadic Na layers is the similarity in the Na layer profiles measured on different days. In Figure 5.11 are plotted Na density profiles measured near 0300 LST on January 18, 21, and 22. The main features of the profiles such as the sharp increase of density near 90 km and slow decrease above 90 km are apparent in all the profiles. The winds induced by the diurnal tide may be responsible for this daily repeated profile shape. Sporadic Na layers developed near 90 km shortly after the profiles of January 18 and 22 were measured. For example, on January 22 the peak at 90 km developed into sporadic layer 5 about 2 min later. Five hours earlier this peak was near 96 km and had developed into the dominant sporadic layer plotted in Figure 5.1c. The peak moved downward at a velocity of 31.4 cm s⁻¹.

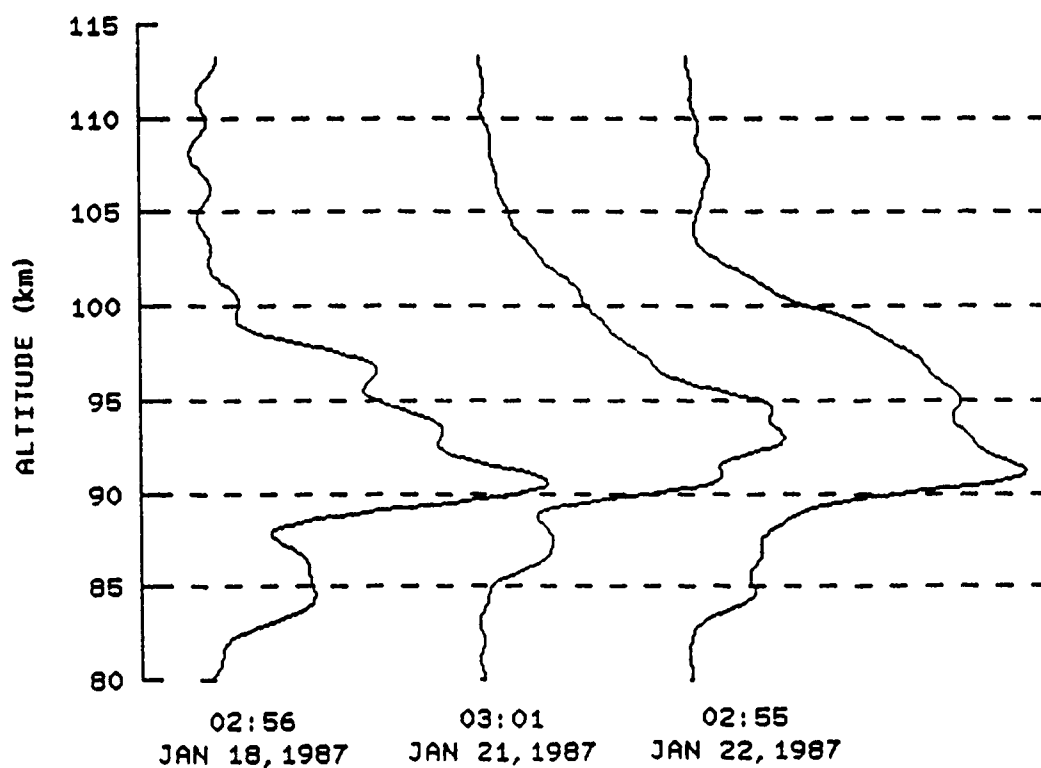


Figure 5.11. Sodium density profiles measured at 0256 LST on January 18, 0301 LST on January 20, and 0255 LST on January 22, 1987.

Although the vertical velocities of the long-lived sporadic layers were comparable with the phase velocity of the diurnal tide, the periods and vertical wavelengths of the waves responsible for the observed velocities could not be determined from the Na lidar data. The altitudes of local density maxima observed on the night of January 21 are plotted in Figure 5.12. The measurement times of the density profiles of Figures 5.1b, 5.1c, 5.2, and 5.4 are marked on the bottom of Figure 5.12. The dominant sporadic layer that started near 95 km at 2200 LST moved steadily downward during the measurement period. Note that other density maxima also moved downward, with similar velocities. The maximum near 91 km, starting at 2200 LST, moved downward for about 2.5 hours, with a velocity of 32 cm s^{-1} ; the maximum near 108 km, starting at 0200 LST, moved downward for about 3 hours, with a velocity of 38 cm s^{-1} . The maximum near 98 km, starting at 0430 LST, also moved downward for about 1.5 hours, with a velocity of 33 cm s^{-1} . If a diurnal tide with a vertical wavelength of 25-30 km was dominating the Na layer dynamics, the Na density maxima created by the tidally induced winds would move downward with the diurnal vertical phase velocities. However, it is not clear how the tidally induced winds would cause the sporadic Na layers to move downward with the same velocity. On the night of January 21, there were at least these four groups of density maxima, including the dominant sporadic layer, moving downward at about $28\text{-}39 \text{ cm s}^{-1}$. The average vertical distance between these groups of the maxima was about 6 km between 2210 and 0050 LST, about 15 km between 0210 and 0500 LST, and about 9 km between 0430 and 0600 LST.

During the formation periods the apparent vertical movements of the sporadic layers also appear to be influenced by the prevailing horizontal winds. During the formation periods of 10 sporadic layers, the upward motions of the sporadic layers and Na density maxima were observed. The maximum velocity of the apparent upward motions was about 1.6 m s^{-1} . Because this upward velocity was quite large and because the upward motions were observed to change into downward motions within a very short period of a few minutes, vertical winds are probably not responsible for the motions. Instead, the advection of the tilted layers across the

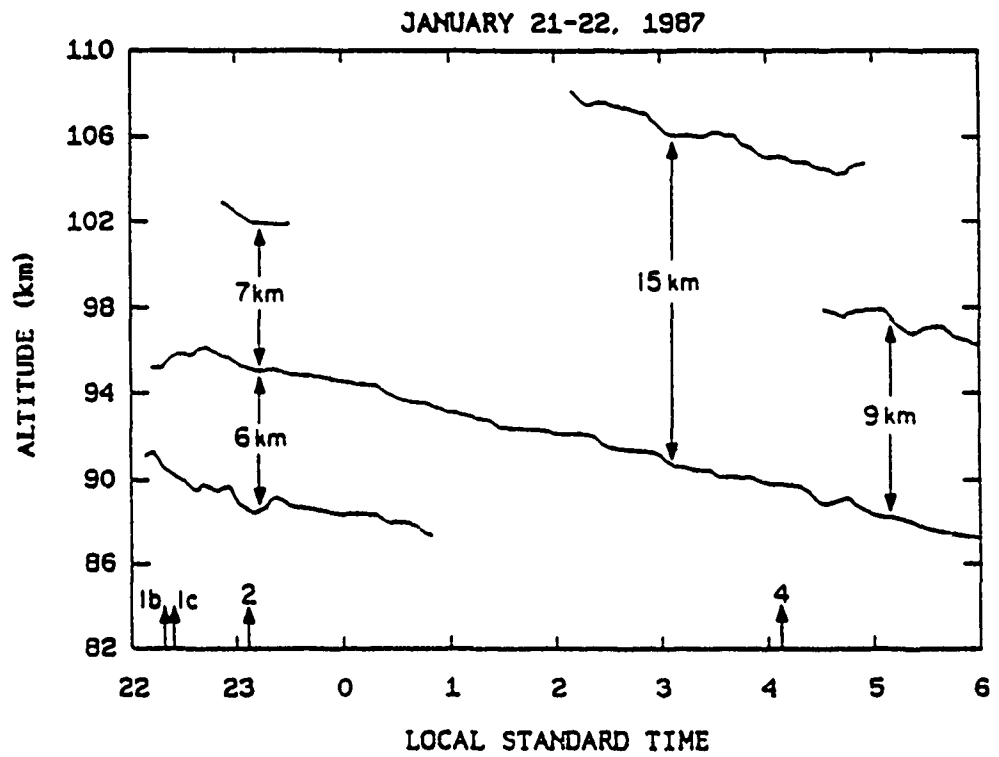


Figure 5.12. Temporal variations of the altitudes of local Na density maxima on January 21-22, 1987 at Mauna Kea Observatory. The measurement times of the density profiles of Figures 5.1b), 5.1c), 5.2, and 5.4 are marked on the bottom of this figure.

lidar site by prevailing horizontal winds might be responsible for the apparent motions. To illustrate, the altitudes of the density maxima and minima observed on the night of January 21 are plotted in detail from 2100 to 2330 LST in Figure 5.13. The measurement times of the density profiles of Figures 5.1a, 5.1b, 5.1c, and 5.2 are marked on the bottom of Figure 5.13. The altitudes of the sporadic layers are represented by solid lines, the altitudes of other density maxima by dashed lines, and the altitudes of density minima by dotted lines. During this period a total of nine sporadic layers were observed. From 2140 to 2223 LST, three sporadic layers (2, 4, and 16) moved downward at an average velocity of 2 m s^{-1} in the altitude region between 96 and 105 km. Then, as the downward movements disappeared, the dominant sporadic layer (1) reached the highest peak density at 2224 LST. From 2226 to 2251 LST, two sporadic layers (9 and 15), along with other density maxima, moved upward at an average velocity of 1.6 m s^{-1} . The dominant layer also moved upward at a velocity of approximately 1.6 m s^{-1} from 2215 to 2229 LST. Between 2229 and 2231 LST the layer was displaced from 96.5 to 95.6 km. Then, from 2231 to 2242 LST the layer moved upward again, at a velocity of approximately 1.5 m s^{-1} . After 2242 LST the layer began moving downward at 31 cm s^{-1} and maintained this velocity for the next 7 hours. Sporadic layer 12 near 103 km began developing at 2251 LST, when the upward motion of the sporadic layer 9 changed to downward motion.

5.4 Summary

The maximum densities at the peaks of the sporadic Na layers observed above Mauna Kea were of the order of 104 cm^{-3} , and the full widths measured at the 80% points were of the order of 1 km. The formation periods of the layers ranged from a few minutes to about 1 hour. Although the lidar campaign was conducted only on 5 nights, so that the data base is not large, the occurrence times of the sporadic layers appear to have a pattern. The layers formed either in the late evening between 2100 and 2330 LST or in the early morning between 0300 and 0600 LST. On 3 different nights four layers began forming within a short time span of 15 min from

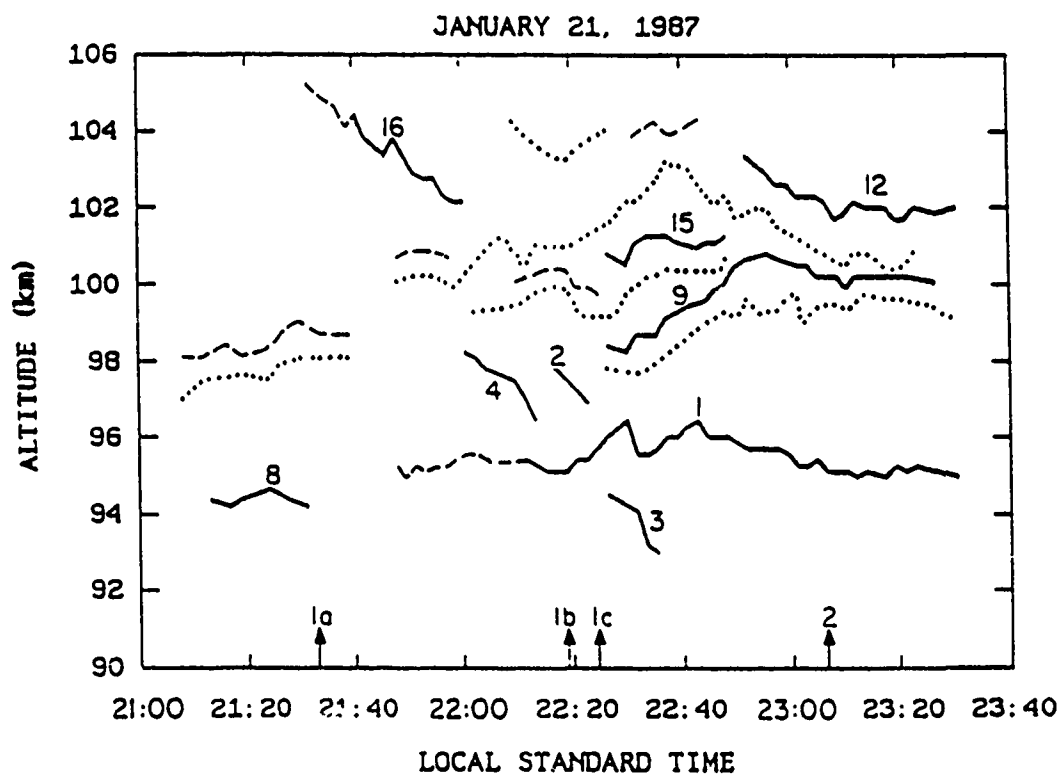


Figure 5.13. Temporal variations of the altitudes of local Na density maxima and minima from 2100 to 2330 LST on January 21, 1987 at Mauna Kea Observatory. The solid curves represent the altitudes of the sporadic layers, the dashed curves the local density maxima, and the dotted curves the local density minima. The sporadic layers are numbered in accordance with Table 5.2. The measurement times of the density profiles of Figures 5.1a), 5.1b), 5.1c), and 5.2 are marked on the bottom of this figure.

0253 to 0308 LST. A total of 10 layers began forming between 2100 and 2300 LST, but no layers formed between 2300 and 0250 LST. The average of the times of the maximum density of the early morning sporadic layers occurred about 6 hours after that of the late evening layers. In general, the formation times of the early morning layers were longer. The starting altitudes of the early morning layers were also lower. In addition to the patterns of occurrence times, the approximately 30 cm s^{-1} apparent downward velocities of the three longest duration layers suggest that the mechanisms responsible for creating the layers may be related to the diurnal tide.

The sporadic Na layer phenomena also appear to be latitude dependent, because the layers have been observed only at low- and high-latitude sites. In fact, at the mid-latitude site of Urbana, Illinois (40°N , 88°W), only one prominent sporadic Na layer has been recently observed during almost 10 years of Na lidar observations. A few minor sporadic Na layers were recently observed at Urbana in March and April 1988, using a new more powerful CEDAR lidar system [Beatty *et al.*, 1988]. However, these layers appear to be meteor trails and were not nearly as spectacular as those observed in Mauna Kea and the Arctic. The characteristics of the sporadic layers observed at the high-latitude site of Andoya and the low-latitude sites of Sao Paulo and Mauna Kea are very similar. Simultaneous occurrences of the sporadic E and Na layers at almost identical altitudes were observed on many occasions at all the three sites. However at Mauna Kea, sporadic E and Na layers were not always observed simultaneously. In fact, for the four sporadic Na layers of the early morning no corresponding sporadic E layers were observed. Increases in Na D-line emission intensities were also observed during the occurrences of sporadic Na layers in Sao Paulo and the high-latitude site of Longyearbyen, Svalbard.

The mechanisms responsible for creating the sporadic Na layers are not well understood. Because of the strong correlation in altitudes and times of the sporadic Na and E layers above Andoya, von Zahn and Hansen [1988] have argued that both are probably created by the same mechanism. The dynamic effects of tides and waves can lead to the development of very thin layers of metallic ions. However, the chemical processes for converting Na^{+} to neutral Na

appear to be too slow to explain the rapid nearly simultaneous development of sporadic E and Na layers. Consequently, *von Zahn and Hansen* [1988] suggest that both layers are formed by the impact of auroral particles on upper atmospheric dust and smoke particles. They argue that this process would evaporate and ionize metals which are absorbed on the surfaces of the dust particles. Clearly, auroral excitation is not responsible for the sporadic Na layers observed above the low-latitude sites of Mauna Kea and Sao Paulo. The starting times of the 16 layers observed at Mauna Kea were restricted to the short 2-hour interval between 2052 and 2251 LST and the 15-min interval between 0253 and 0308 LST. It is interesting that the 10 sporadic Na layers observed on 7 different nights and analyzed by *von Zahn and Hansen* [1988] also had starting times restricted to the short 3-hour interval between 2159 and 0043 LST and had formation periods which were comparable to those of the late evening layers above Mauna Kea. It is difficult to see how auroral excitation would lead to such restricted occurrence times. It is also puzzling that sporadic Na layers appear to be very rare at mid-latitudes, while sporadic E is not. Clearly, additional observations, particularly at low latitudes, are needed to better characterize this very interesting phenomenon.

6. LIDAR OBSERVATIONS OF THE SODIUM LAYER AT SVALBARD, NORWAY

6.1 Overviews

The UIUC group conducted a total of five Na lidar campaigns at Nordlysstasjonen (78°12'N, 15°50'E), Svalbard, Norway from July, 1987 to April, 1988. Table 6.1 summarizes the observation periods of the five campaigns. The seasonal variations of the Na column abundance measured at Svalbard and Urbana are compared in Figure 6.1. Symbols denote average values for the observation period, while lines denote range of values. The dashed lines indicate the abundance measured at Andoya, Norway, which is located approximately 1000 km south of Svalbard. The Na abundance measured at Urbana shows a distinct annual oscillation with a sharp peak in November to January and a broader minimum in the summer months. This annual oscillation is believed to be related to changes in the mesopause temperature that affect the reaction rates of the main chemical loss processes for Na [Swider, 1985; Jegou *et al.*, 1985b]. The seasonal variations of the abundance measured at Svalbard also show an annual oscillation with a broad maximum in January to April followed by a distinct minimum in June. The minimum abundance observed in June is not well understood, and will be discussed further in Section 6.2. The maximum abundance observed in January to April at Svalbard is comparable to the maximum abundance in November to January observed at Urbana.

The seasonal variations of the layer centroid height measured at Svalbard and Urbana are compared in Figure 6.2. The centroid measured at Urbana shows a semi-annual oscillation with maxima near equinox in March and September and minima near solstice in June and December. The centroid measured at Svalbard also shows the similar semi-annual oscillations with maxima near equinox in March and September and minima near solstice in December and June. During the winter months, the centroid height observed at Svalbard was generally lower than the

Table 6.1. Sodium Lidar Observation Times at Svalbard (1987-1988)

Campaign	Start		End	
	Date	Time(UT)	Date	Time(UT)
#1	July 10	2200	July 11	2255
	July 12	0540	July 12	1045
	July 12	1755	July 12	1835
	July 13	1330	July 14	0005
	July 14	1825	July 14	2320
#2	September 7	0800	September 7	1150
	September 7	2055	September 7	2335
	September 8	0310	September 9	1215
	September 10	1725	September 11	0045
#3	October 31	1600	October 31	1615
	November 7	1340	November 9	2230
	November 11	1855	November 11	2315
	November 13	0800	November 13	0930
#4	January 6	1525	January 6	2225
	January 7	0120	January 9	0110
	January 9	2300	January 13	1350
	January 14	1750	January 14	2035
	January 15	0825	January 15	1225
	January 22	0800	January 22	2200
	January 24	1650	January 24	2150
#5	April 4	2030	April 5	0130
	April 5	2335	April 6	0000
	April 6	2045	April 7	0140
	April 8	0110	April 9	0010
	April 10	2150	April 12	0445
	April 12	1855	April 13	0645
	April 13	1735	April 14	0555
	April 18	1800	April 18	2035
	April 19	0055	April 19	0525

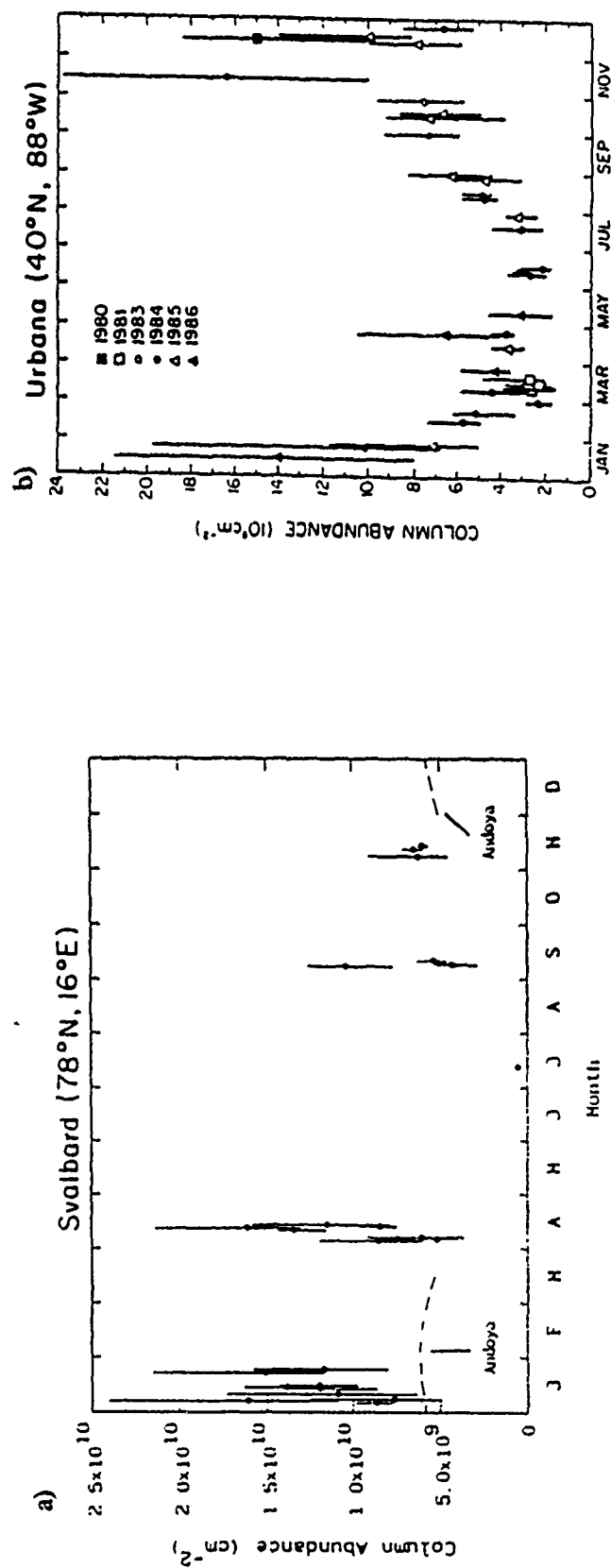


Figure 6.1 Seasonal variations of Na column abundance measured a) at Svalbard, Norway (78°N) and b) at Urbana, Illinois (40°N). Symbols denote average values for the observation period, while lines denote range of values.

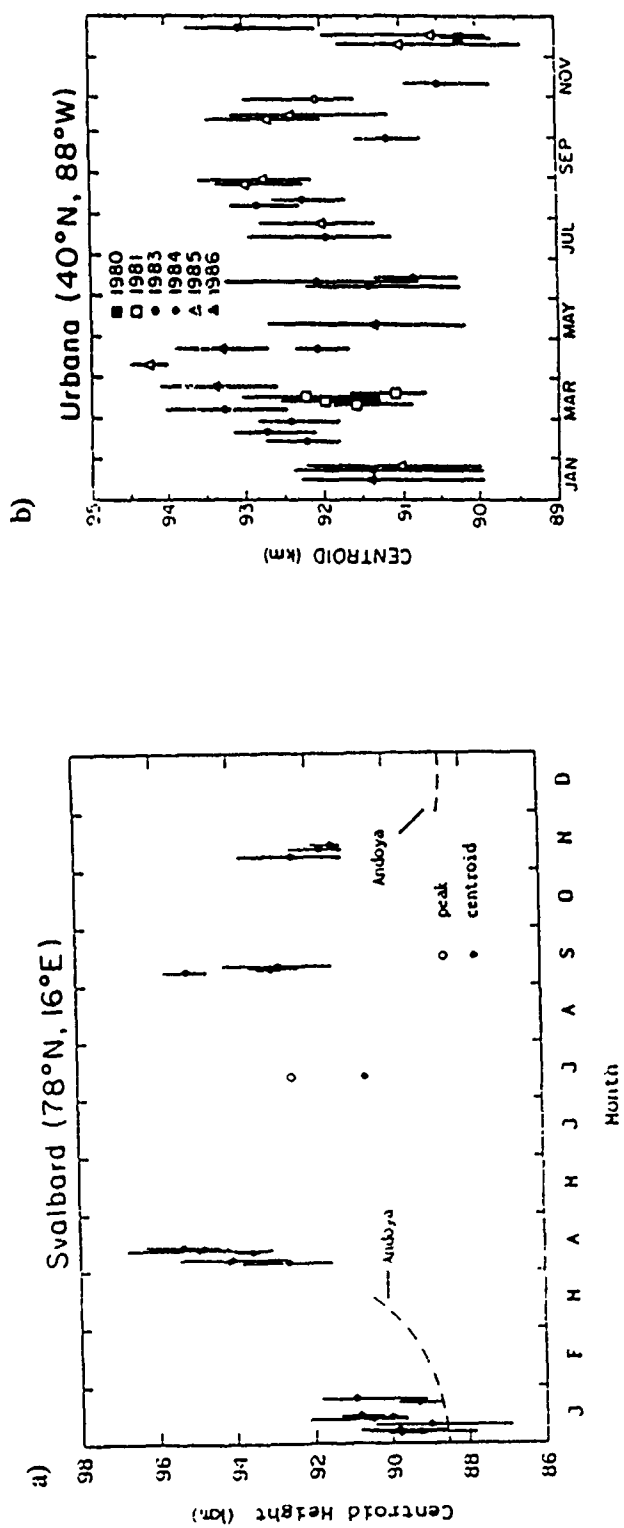


Figure 6.2. Seasonal variations of Na layer centroid height measured a) at Svalbard, Norway (78°N) and b) at Urbana, Illinois (40°N). Symbols denote average values for the observation period, while lines denote range of values.

centroid observed at Urbana. This could be the result of a stronger downward motion of the atmosphere over the high latitude site of Svalbard during the winter months.

In Figure 6.3, the seasonal variations of the layer rms width measured at Svalbard and Urbana are plotted. The rms width measured at Urbana exhibits no clear seasonal variations. However, the rms width measured at Svalbard shows an annual oscillation with a maximum near February and a minimum near September. The mechanisms responsible for this annual oscillation are not well understood.

Sporadic Na layers were observed at Svalbard during every campaign except the July 1987 campaign. The characteristics of these sporadic Na layers are quite similar to those observed at Andoya, Norway by *von Zahn and Hansen* [1988] and at Mauna Kea Observatory, Hawaii by *Kwon et al.* [1988]. The characteristics of the Na layer observed during the July and September 1987 campaigns will be presented in Section 6.2, and the characteristics observed during the November 1987 and January 1988 campaigns will be presented in Section 6.3.

6.2 Lidar Observations at Svalbard in July and September, 1986

The chemistry and dynamics of the mesospheric Na layer have been studied extensively since the late 1960s with lidar techniques. Meteoric ablation is generally regarded as the dominant source of all mesospheric alkali metals including Na. The Na layer is typically confined to the region between 80 and 110 km with a peak near the mesopause at 90 km, where the density ranges from about 10^3 to 10^4 cm⁻³. The Na column abundance at mid-latitudes in the Northern Hemisphere varies from a summer minimum of about 3×10^9 cm⁻² to a winter maximum of about 10^{10} cm⁻² in December and January [*Gardner et al.*, 1986]. The seasonal and geographical variations in Na abundance are now believed to be related to changes in the mesopause temperature that affect the reaction rates of the main chemical loss processes for Na. In addition to chemical activity, the dynamic effects of the tides, gravity waves and the mean winds have a significant influence on the vertical structure of the layer.

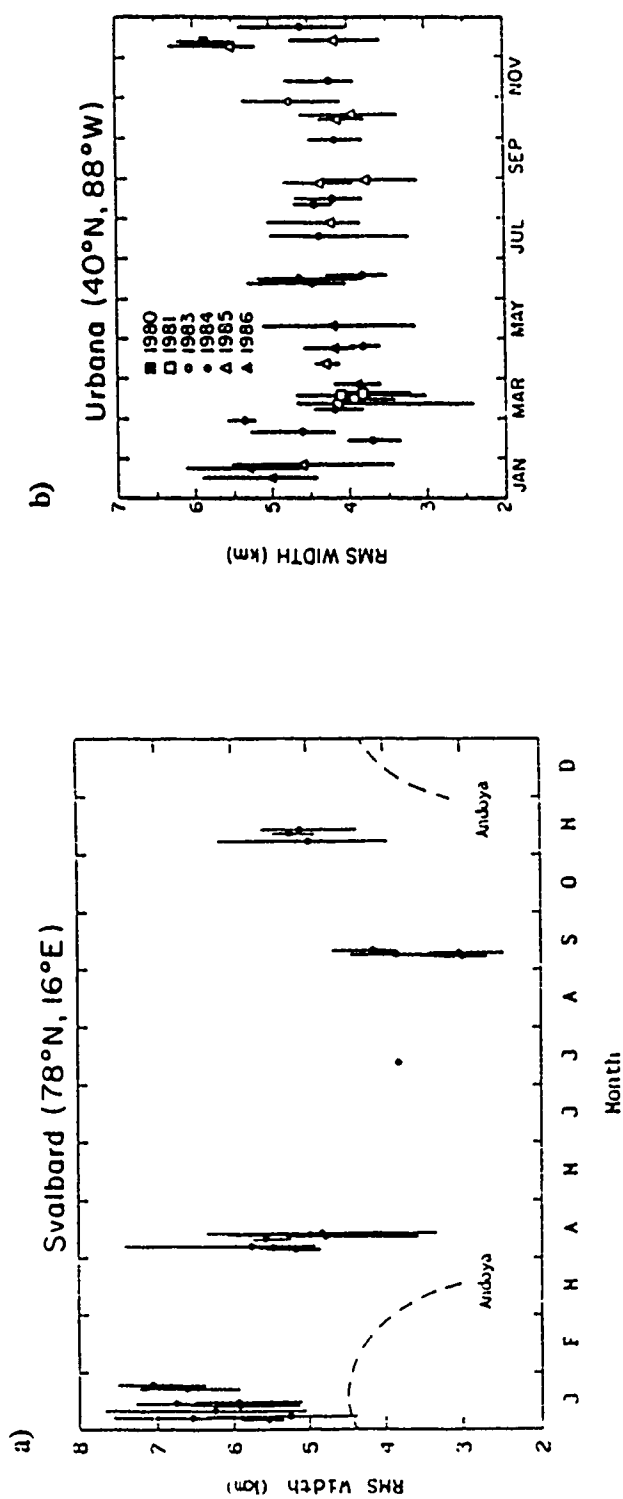


Figure 6.3. Seasonal variations of Na layer rms width measured a) at Svalbard, Norway (78°N) and b) at Urbana, Illinois (40°N). Symbols denote average values for the observation period, while lines denote range of values.

The system parameters and daytime capabilities of the UIUC Na lidar are described by *Kwon et al.* [1987]. Data processing and analysis are discussed by *Gardner et al.* [1986]. Figure 6.4 is a plot of the average Na profile measured during the period July 10-15, 1987 at Nordlysstasjonen, Svalbard. Figures 6.5 and 6.6 are plots of the profiles measured near local midnight on September 7 and 9, 1987, respectively. Table 6.2 lists the major Na layer parameters for these three profiles. The large standard deviations in the parameters measured during July are caused by the low Na signal level and the high background noise from the bright sunlit sky.

Although in July the centroid height and rms thickness are comparable to values measured at other locations, the peak density and column abundance are almost a factor of 5 lower. These results seem to be the first Na measurements made near the North Pole during summer, are very surprising, particularly when compared with other high latitude observations. It is well-known that Na abundance is quite high in the wintertime polar mesosphere. Lidar measurements by *Megie et al.* [1978] and *Juramy et al.* [1981] at Heyss Island, Franz Joseph Land (80°N, 50°E), by *von Zahn and Tilgner* [1987] at Andoya, Norway (69°N, 16°E) and by *Nomura et al.* [1987] at Syowa Station, Antarctica (69°S, 40°E) show Na abundances ranging from 3×10^9 to $12 \times 10^9 \text{ cm}^{-2}$ during the polar winter. In addition, *von Zahn et al.* [1988] report that Na abundance in July at Andoya averaged approximately $6 \times 10^8 \text{ cm}^{-2}$, which is consistent with the July measurements at Nordlysstasjonen approximately 1000 km to the north.

The September 7 profile plotted in Figure 6.5 is also quite interesting because of the high peak density and small thickness of the layer. The column abundance for this profile is almost 24 times larger than the abundance measured in July. The 2 km half-width and high peak density are characteristic of the very narrow layers which *von Zahn et al.* [1988] have observed on occasion forming rapidly above Andoya. The profile plotted in Figure 6.6 is more typical of the layer structure observed at Nordlysstasjonen during the September campaign. The column

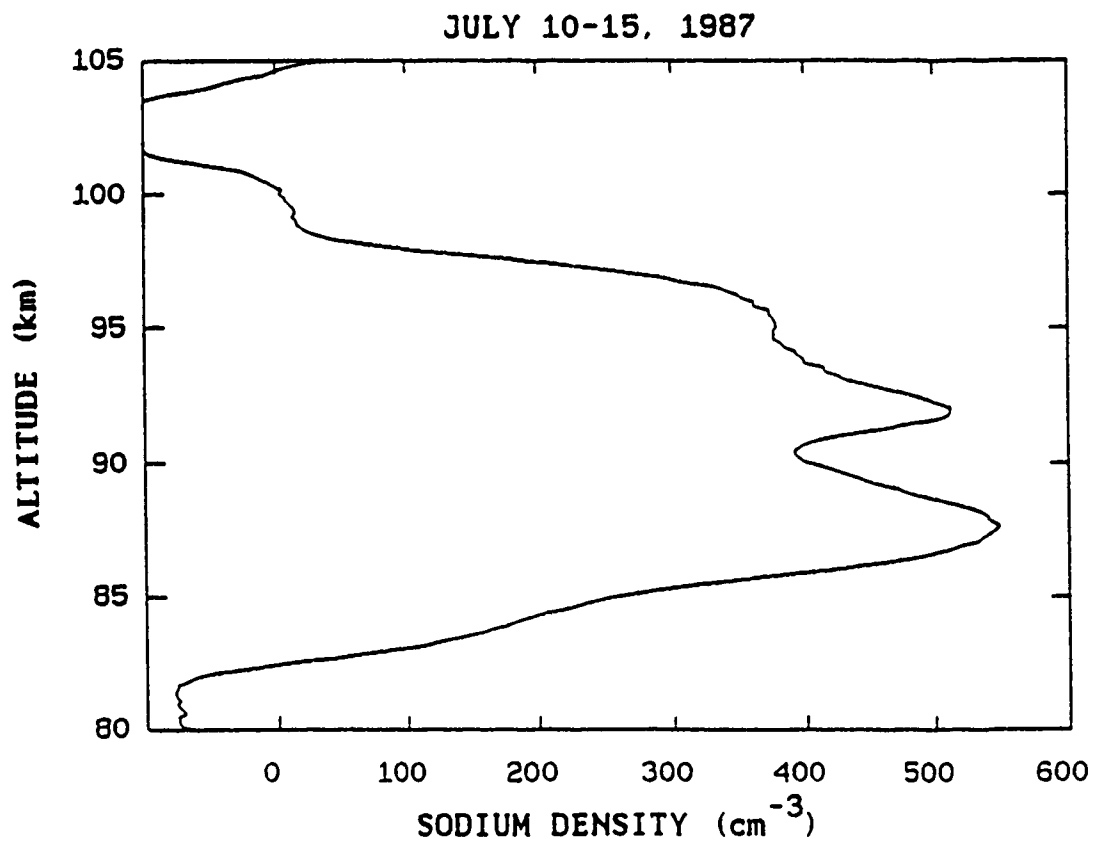


Figure 6.4. Average Na density profile above Nordlysstasjonen, Svalbard during the period July 10-15, 1987. The data were smoothed using modified Hamming window with FWHM = 6 km.

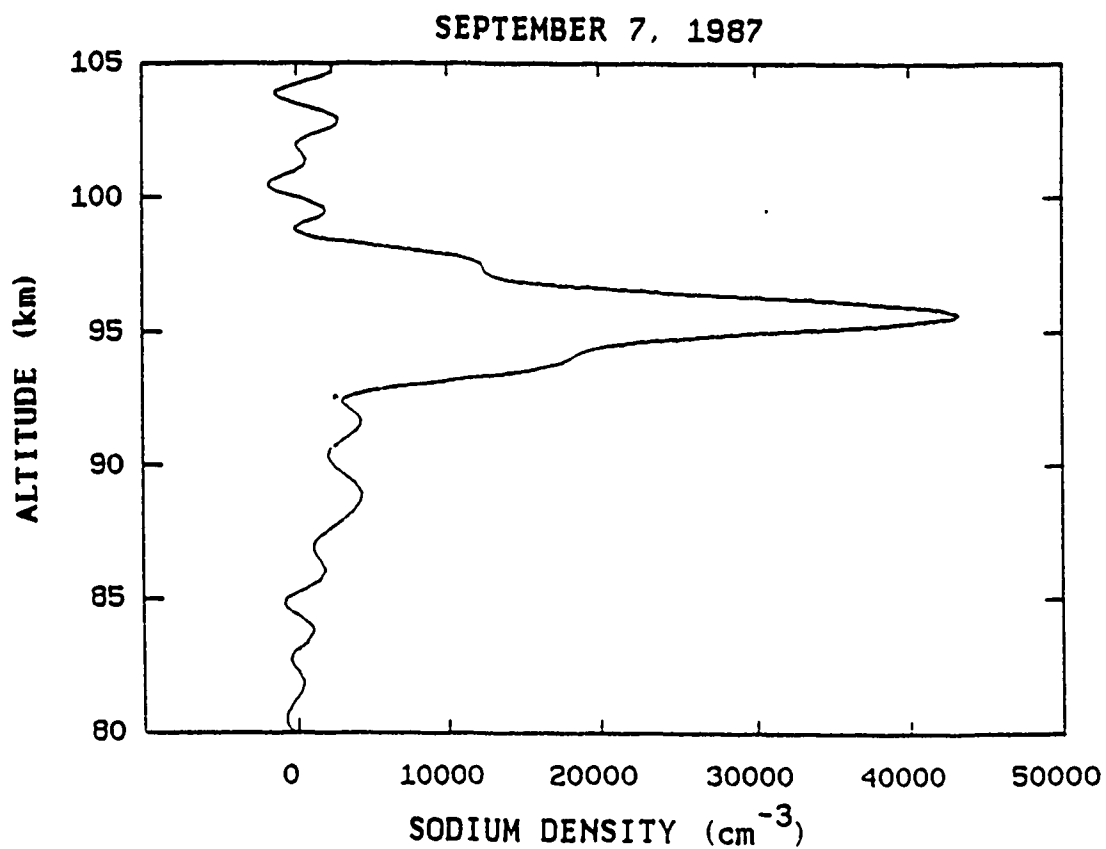


Figure 6.5. Sodium density profile above Nordlysstasjon, Svalbard at 2152 UT September 7, 1987. The integration period was 14 min, and the data were spatially smoothed using a low-pass filter with a cutoff frequency of 0.6 km^{-1} .

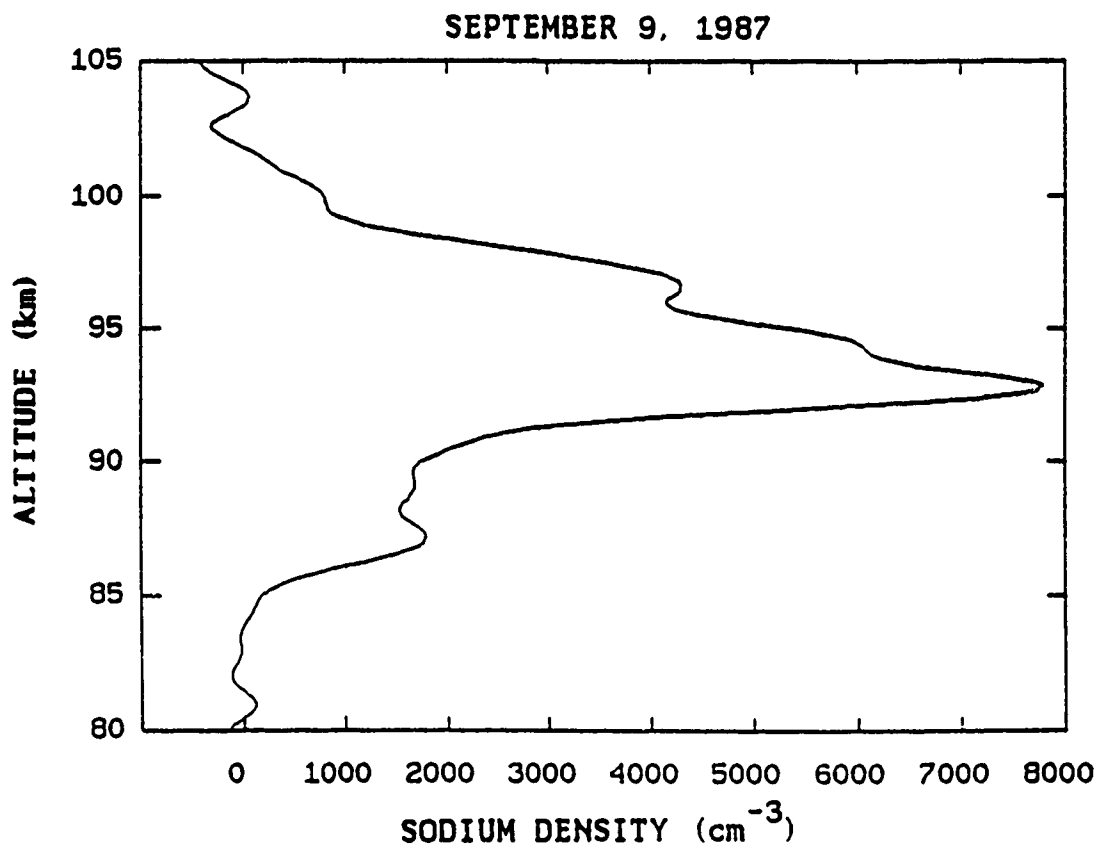


Figure 6.6. Sodium density profile above Nordlysstasjonen, Svalbard at midnight September 9, 1987. The integration period was 40 min, and the data were spatially smoothed using a low-pass filter with a cutoff frequency of 0.6 km^{-1} .

Table 6.2. Sodium Layer Parameters Measured at Nordlysstasjonen (78°12'N, 15°15'E), Svalbard During the July and September 1987 Campaigns

Observation Period	2159 UT July 10 - 0100 UT July 15, 1987	2152-2206 UT September 7, 1987	0000-0400 UT September 9, 1987
Peak Density (cm ⁻³)	550 ± 50	(4.3 ± 0.2)x10 ⁴	(7.8 ± 0.4)x10 ³
Column Abundance(cm ⁻²)	(5.7 ± 2)x10 ⁸	(1.4 ± 0.1)x10 ¹⁰	(4.8 ± 0.1)x10 ⁹
Centroid Height (km)	90.6 ± 2	94.6 ± 0.2	93.5 ± 0.1
RMS Thickness (km)	3.82 ± 2	2.52 ± 0.3	3.23 ± 0.1

abundance ($\sim 4.8 \times 10^9 \text{ cm}^{-2}$) is comparable to mid-latitude values that have been measured in the Northern Hemisphere in early September [Gardner *et al.*, 1986].

The chemistry of the Na layer is still poorly understood even though numerous models have been proposed in an attempt to explain the general feature of the seasonal, diurnal and geographical variations in the layer. Models for the dominant loss mechanisms have generally evolved in two directions. One group of models has been developed by assuming neutral reactions dominate Na chemistry [Kirchhoff, 1986; Thomas *et al.*, 1983]. The other group assumes that ionic species, especially cluster ions of the form $\text{Na}^+ (\text{H}_2\text{O})_n$, are of considerable importance [Richter and Sechrist, 1979; Jegou *et al.*, 1985b]. A complete understanding of the Na layer chemistry has been hampered by the lack of reliable estimates of the rate coefficients for many of the important reactions. Even so, it is generally agreed that the seasonal variation in mesopause temperature is primarily responsible for the seasonal variation of Na abundance at mid-latitudes.

Swider [1985] recently reported model calculations which suggest that the neutral reaction $\text{Na} + \text{O}_2 + \text{M} \rightarrow \text{NaO}_2 + \text{M}$, is an important sink for mesospheric Na. The Na loss rate due to this reaction has a T^{-4} temperature dependence. Von Zahn and Neuber [1987] report that the winter mean temperature between 80 and 100 km at Andoya (69°N) ranges from ~ 200 to 220°K . In early August 1982, Philbrick *et al.* [1984] measured temperatures for the same altitude region at Kiruna, Sweden (68°N). Their values were in the range from ~ 100 to 140°K . Thus the Na loss rate at high latitudes for the neutral reaction discussed by Swider could be as much as 10 times larger in summer compared to winter. However, because the July Na abundance at Nordlysstasjonen (78°N) is almost 24 times smaller than the September 7 levels and almost 9 times smaller than the September 9 levels, this process does not appear to be completely responsible for the observed summertime depletion above Svalbard. Other loss mechanisms which may play significant roles in Na depletion at high latitudes include absorption by polar mesospheric cloud particles, absorption on smoke or dust particles and photo-ionization.

To summarize, the Na abundance above Nordlysstasjonen, Svalbard during July 10-15, 1987 was observed to be almost 5 times lower than typical summertime values measured at mid-latitudes and from 9 to 24 times lower than values measured at Nordlysstasjonen during September 7-11, 1987. Although the very cold temperatures near the summertime Arctic mesopause will increase the effectiveness of the dominant chemical loss process for Na, this mechanism does not appear to be strong enough to explain the large depletion.

6.3 Lidar Observations at Svalbard in November, 1987 and January, 1988

During the November 1987 campaign, the longest observation was made for 57 hours starting at 1500 LST on November 7. The temporal variations and temporal power spectra of the Na column abundance, centroid height, and rms width measured during this observation are plotted in Figures 6.7 through 6.9. The periods of the most dominant variations in the layer parameters were approximately 24 hours and 8 hours. The 8-hour period oscillations were also dominant in the vertical wind velocity estimated at the altitude of 98 km plotted in Figure 6.10. The technique for estimating the vertical wind velocity was described in detail by *Kwon et al.* [1987].

During the January 1988 campaign, the longest observation was made for 72 hours starting at local midnight on January 10. Figures 6.11 through 6.13 show the temporal variations and temporal power spectra of the column abundance, centroid height, and rms width measured during this observation. The periods of the dominant oscillations in the abundance and rms width are 38, 18, and 10 hours, and those in the centroid height are 22, 13, and 8 hours. The 8-hour period oscillations were also dominant in the temporal variations of the vertical wind velocity estimated at the altitude of 98 km as plotted in Figure 6.14. It appears that the Na layer observed in November, 1987 and January, 1988 was dominated by waves with periods of approximately 24 hours and 8 hours.

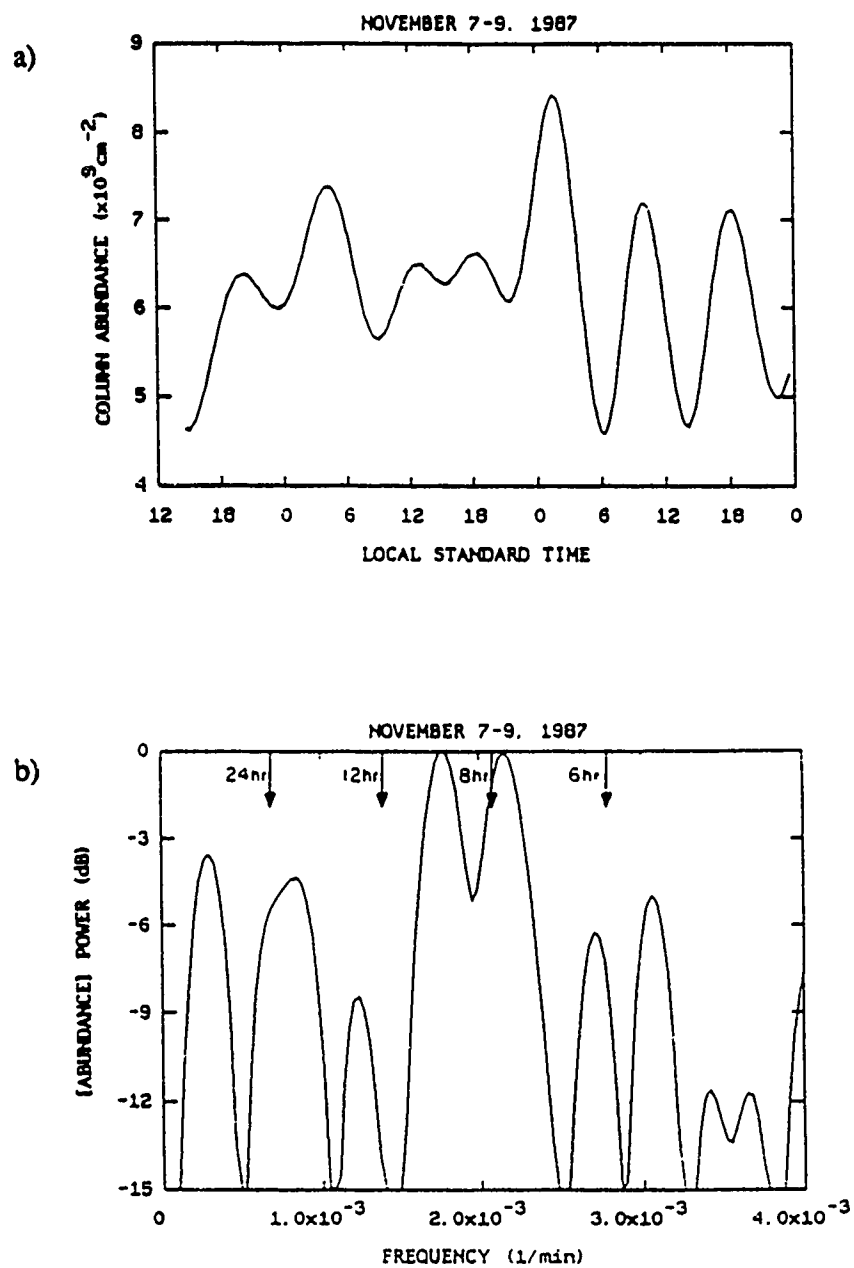


Figure 6.7. a) Temporal variations and b) temporal power spectrum of the Na column abundance measured during the 57-hour period starting at 1500 LST on November 7, 1987.

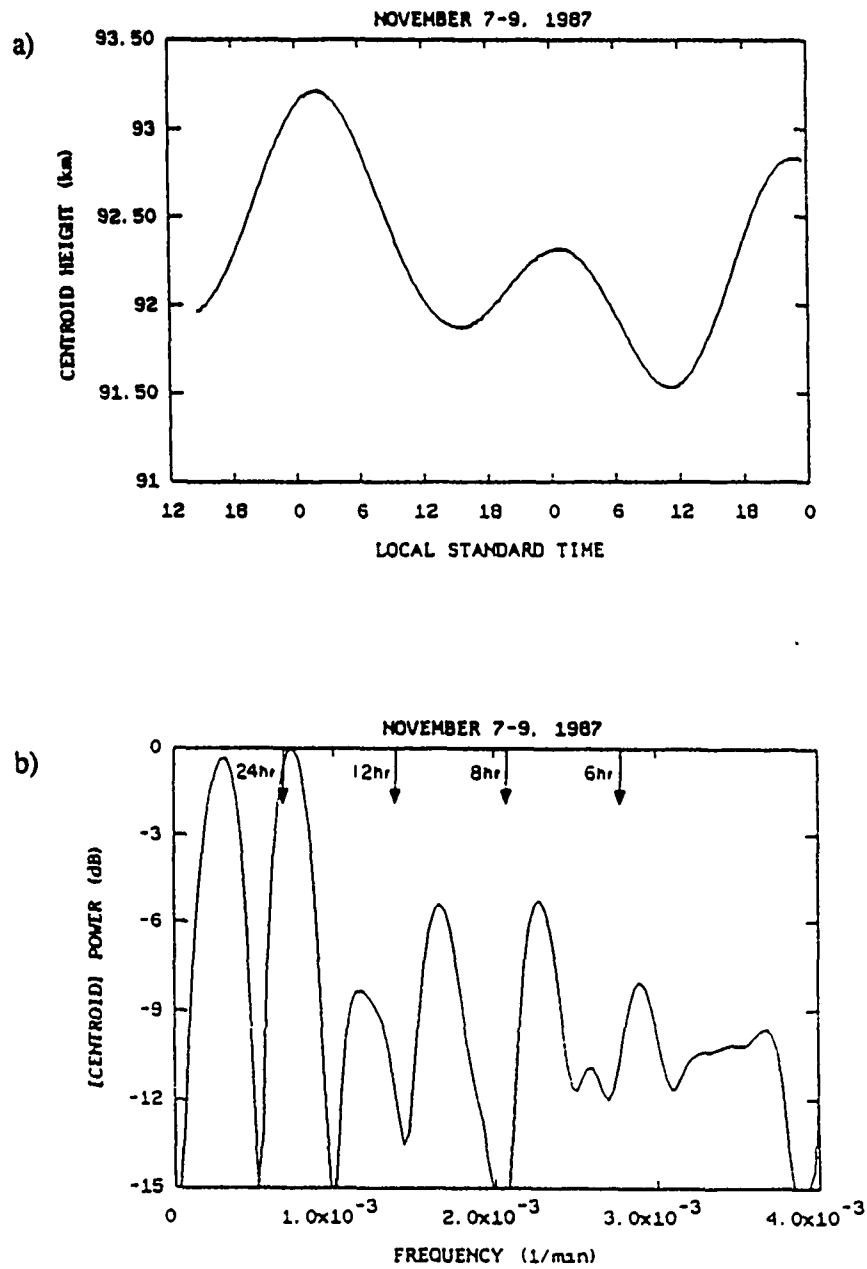


Figure 6.8. a) Temporal variations and b) temporal power spectrum of the Na layer centroid height measured during the 57-hour period starting at 1500 LST on November 7, 1987.

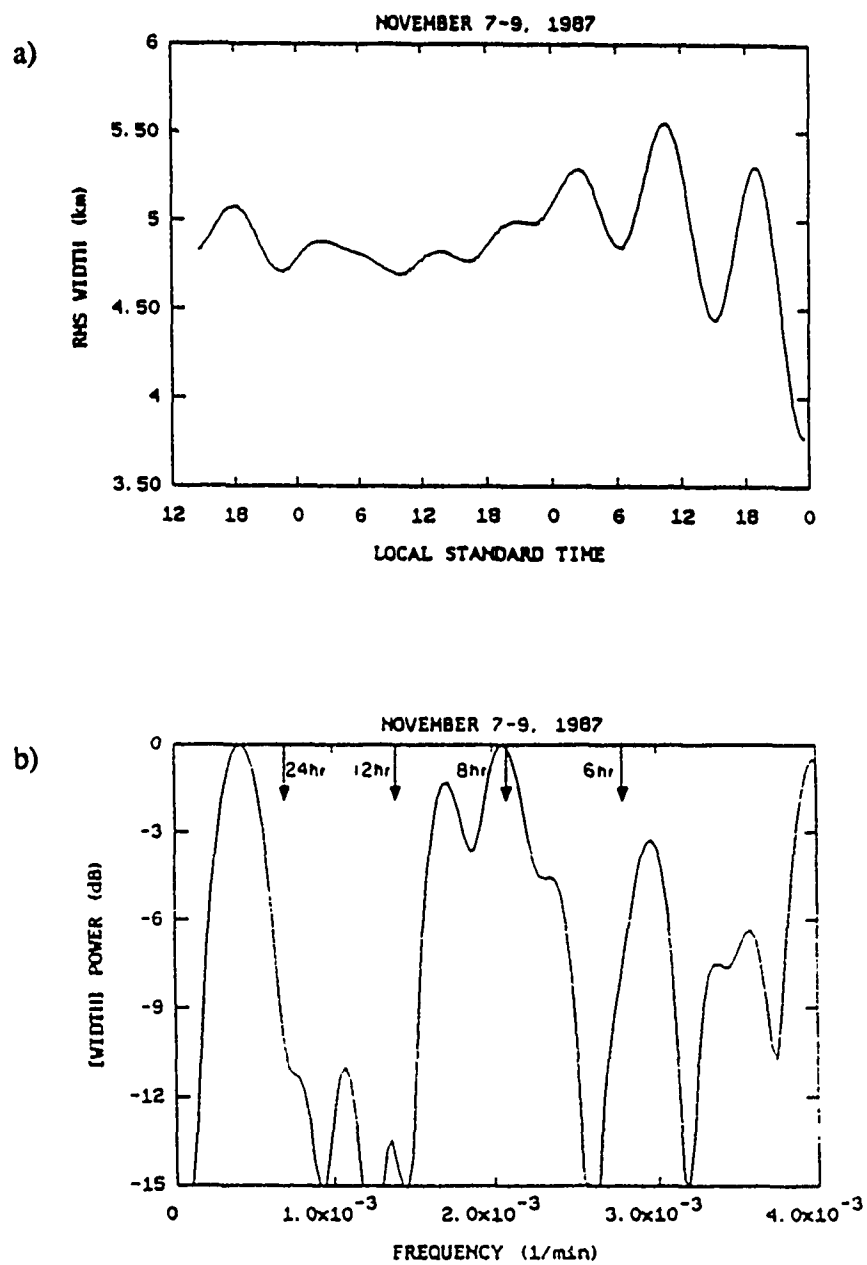


Figure 6.9. a) Temporal variations and b) temporal power spectrum of the Na rms width measured during the 57-hour period starting at 1500 LST on November 7, 1987.

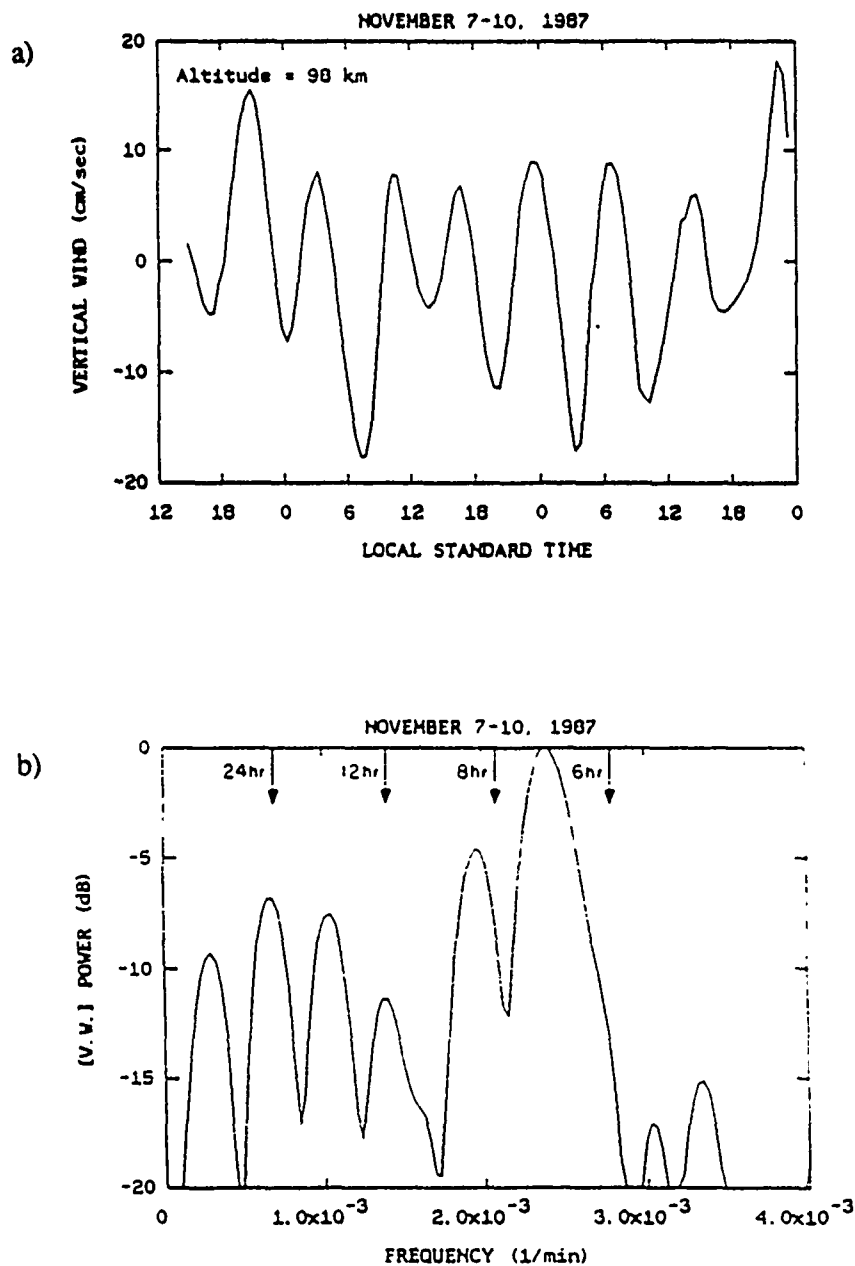


Figure 6.10. a) Temporal variations and b) temporal power spectrum of the vertical winds at the altitude of 98 km. The vertical winds were inferred from the temporal variations of the Na density gradients on the layer topside measured during the 57-hour period starting at 1500 LST on November 7, 1987.

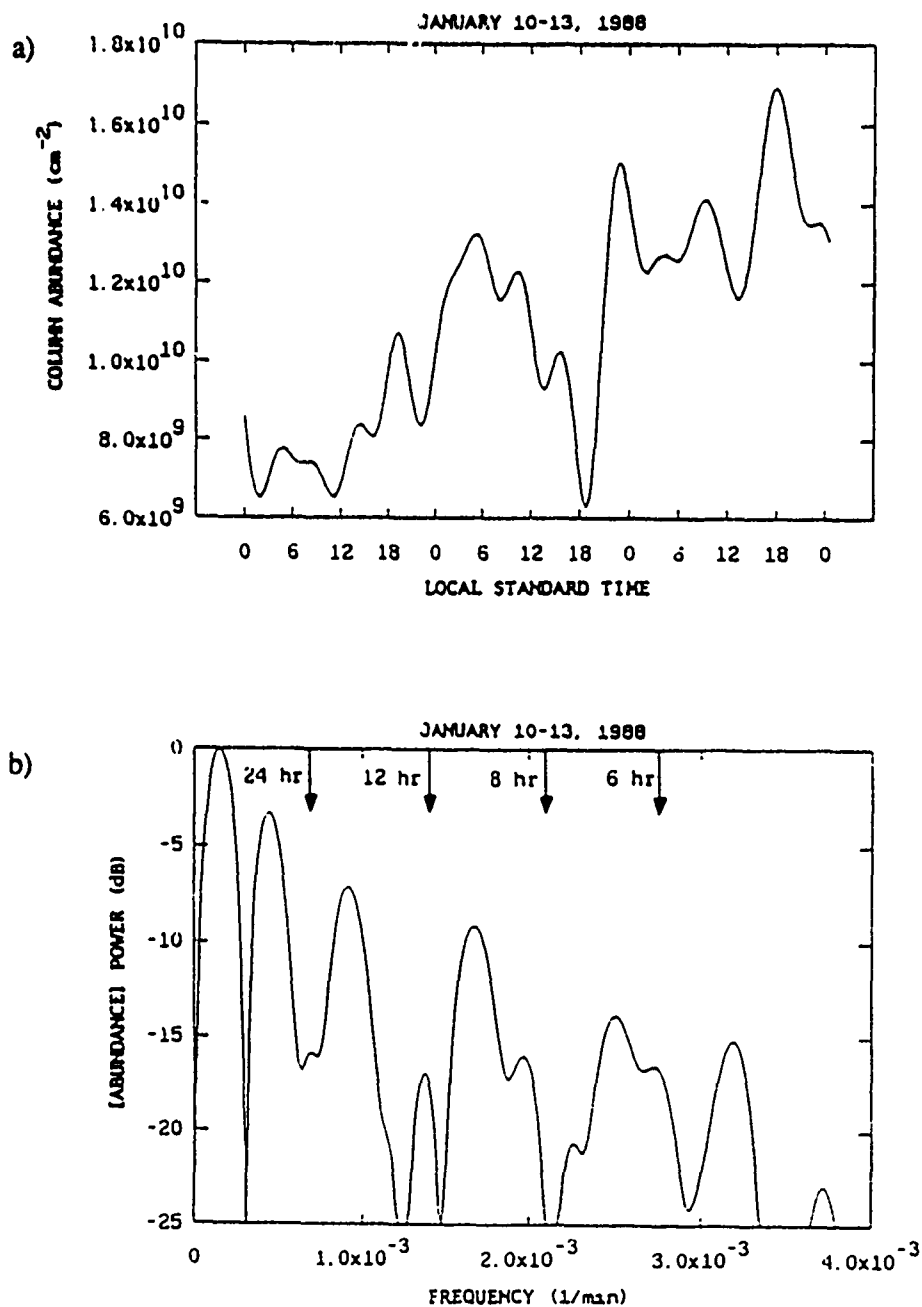


Figure 6.11. a) Temporal variations and b) temporal power spectrum of the Na column abundance measured during the 72-hour period starting at local midnight on January 10, 1988.

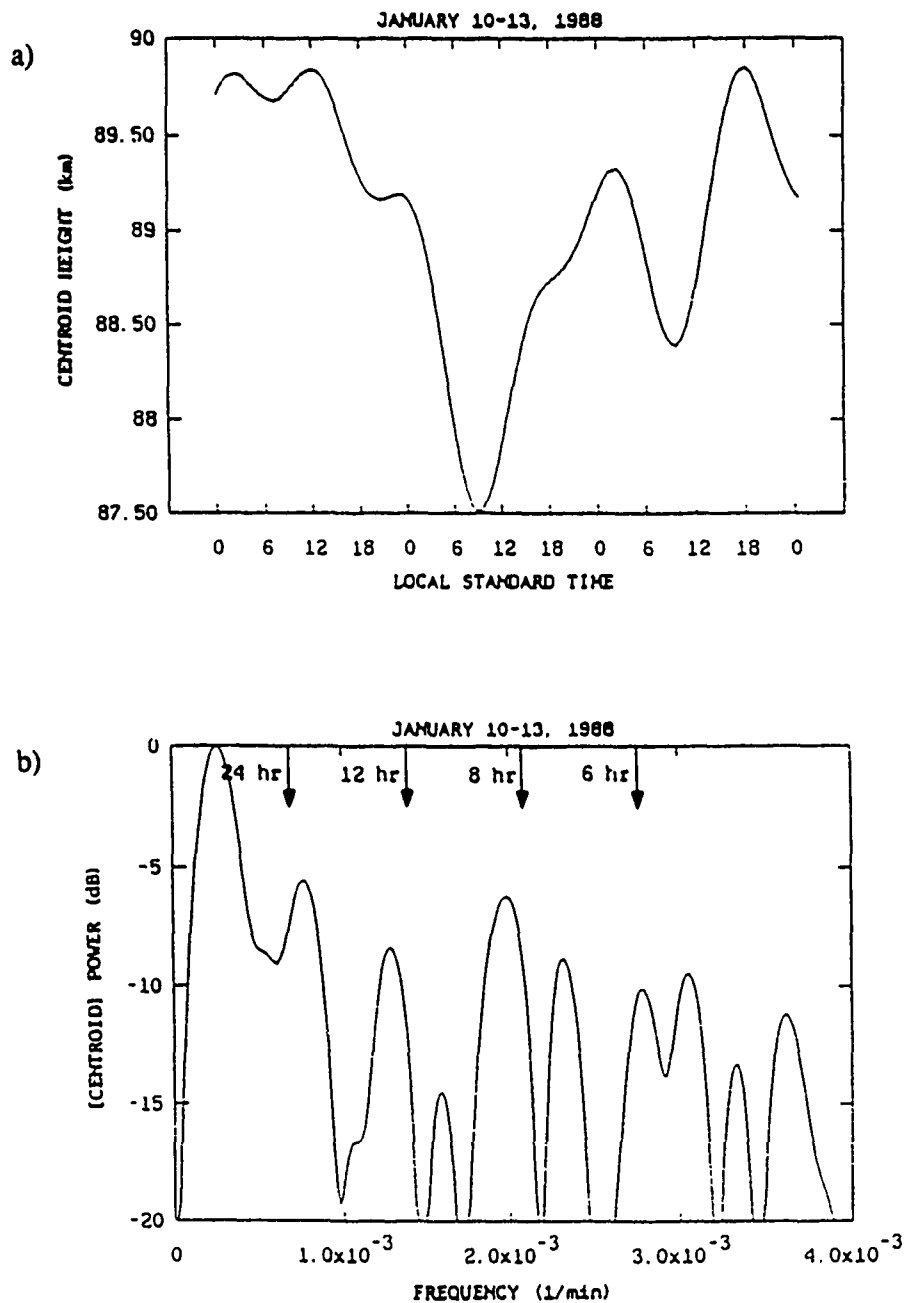


Figure 6.12. a) Temporal variations and b) temporal power spectrum of the Na centroid height measured during the 72-hour period starting at local midnight on January 10, 1988.

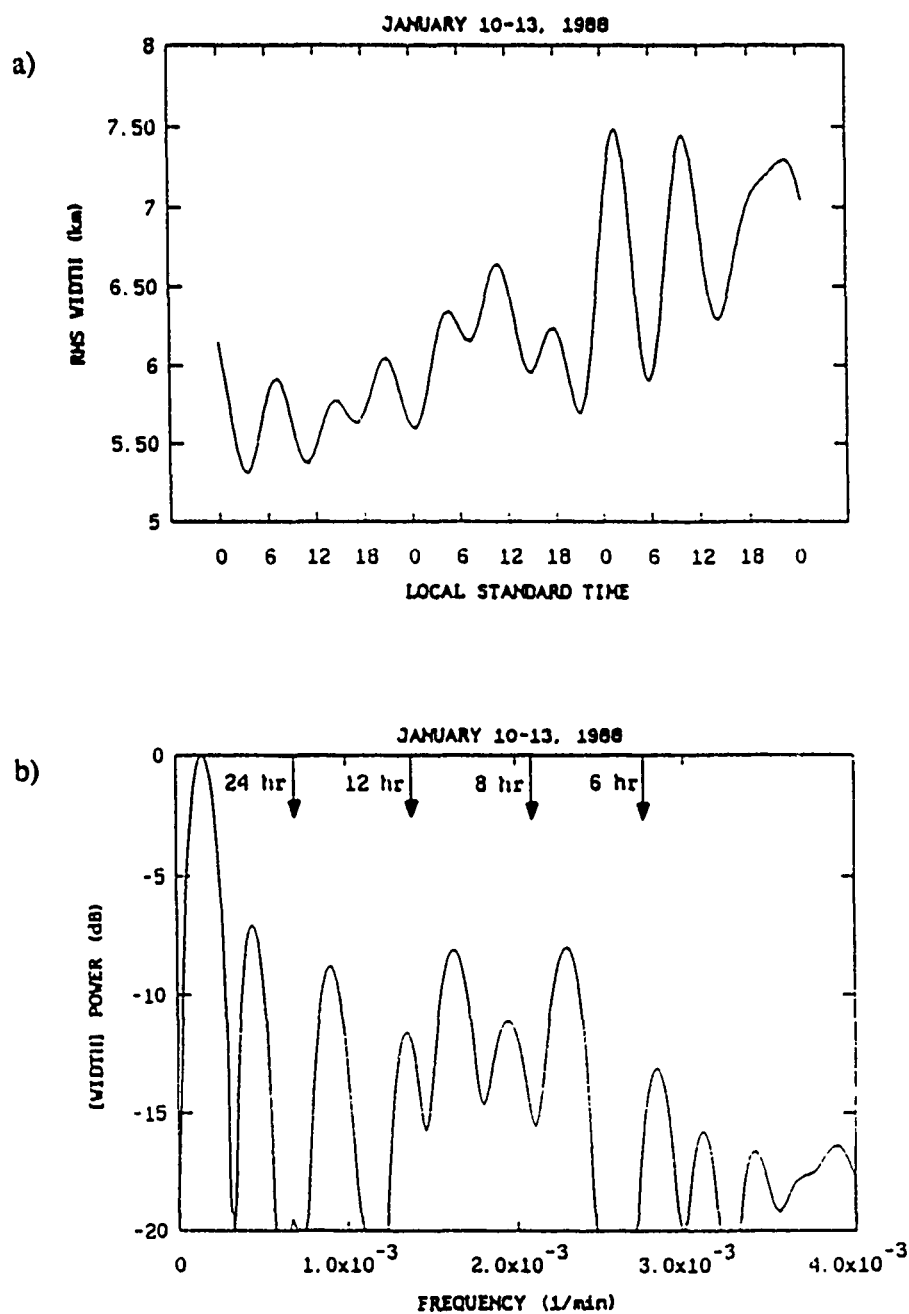


Figure 6.13. a) Temporal variations and b) temporal power spectrum of the Na rms width measured during the 72-hour period starting at local midnight on January 10, 1988.

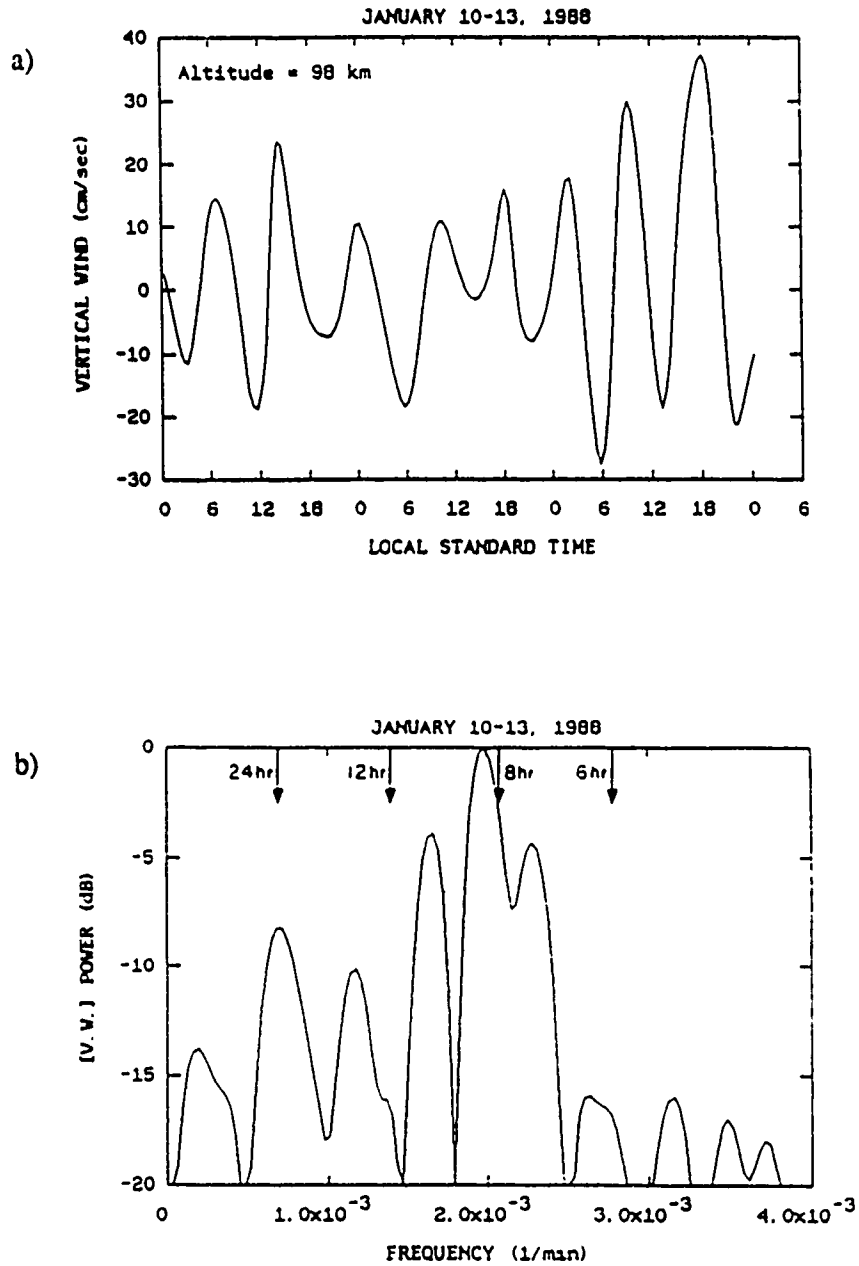


Figure 6.14. a) Temporal variations and b) temporal power spectrum of the vertical winds at the altitude of 98 km. The vertical winds were inferred from the temporal variations of the Na density gradients on the layer topside measured during the 72-hour period starting at local midnight on January 10, 1988.

The characteristics of the sporadic Na layers observed at Svalbard are very similar to those observed at Andoya, Norway by *von Zahn and Hansen* [1988] and at Hawaii by *Kwon et al.* [1988]. To illustrate, sporadic Na layers observed during the November 1987 and January 1988 campaigns are plotted in Figure 6.15. The peak densities and thicknesses of these sporadic Na layers are quite comparable to those observed at Andoya and Hawaii. The sporadic Na layer observed on January 10 developed at the altitude of about 85 km, which appears to be the lowest altitude of the sporadic Na layers observed with the UTUC lidar. Another sporadic Na layer observed during the November 1987 campaign is plotted in Figure 6.16. The temporal evolution of the bifurcation of this sporadic Na layer is plotted in Figure 6.16 b.

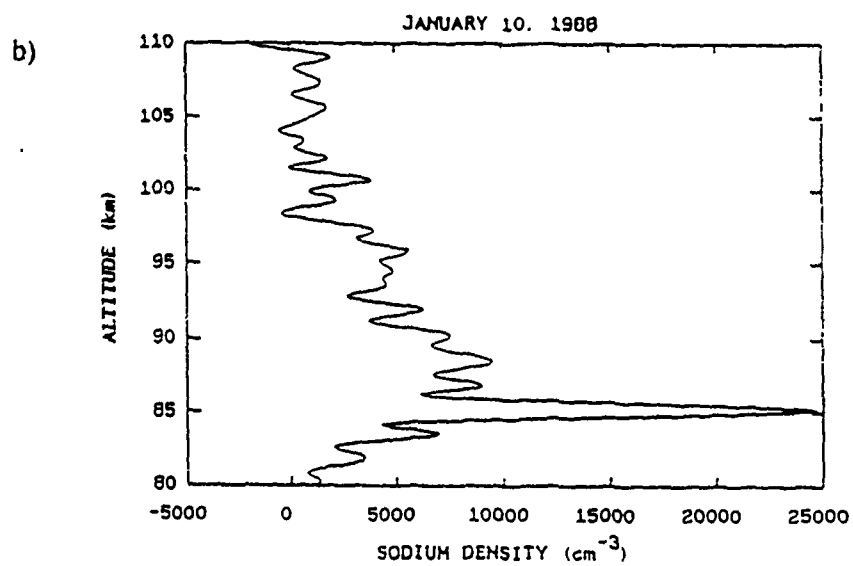
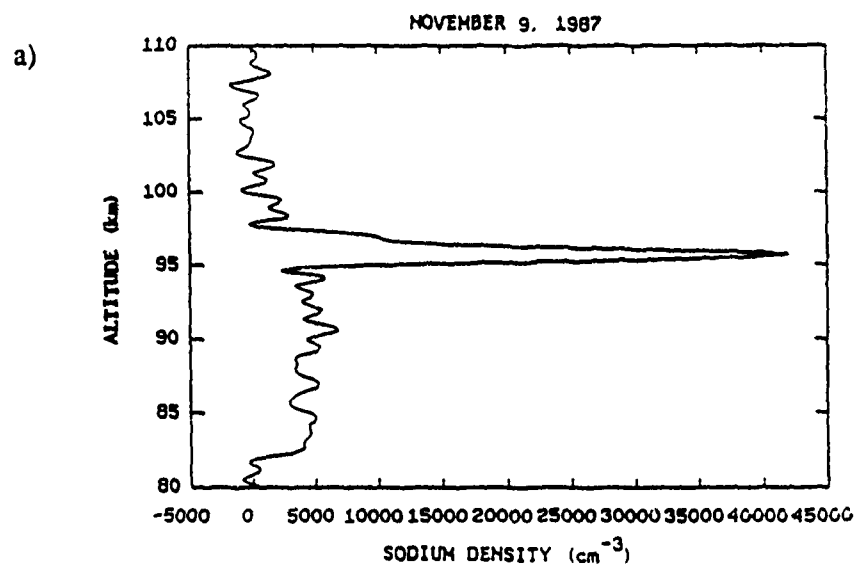


Figure 6.15. Examples of sporadic Na layers observed at Svalbard a) on November 9, 1987 and b) on January 10, 1988.

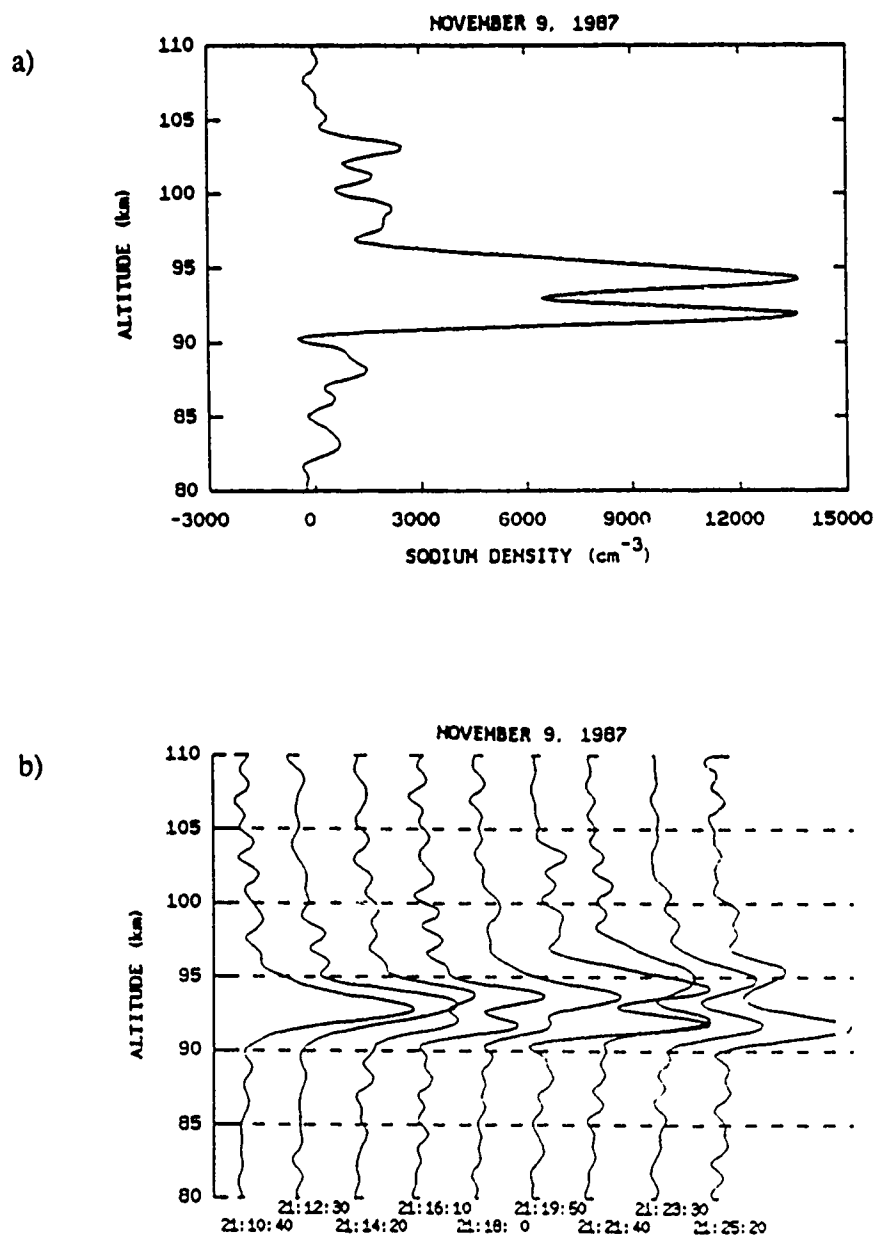


Figure 6.16. a) A bifurcated sporadic Na layer observed at Svalbard on November 9, 1988, and b) the temporal evolution of the sporadic Na layer.

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

A data analysis technique for determining gravity wave intrinsic parameters is described in Chapter 2. The intrinsic parameters include the horizontal and vertical wavelengths, period, and horizontal propagation direction. The technique involves measuring the altitude variations of the wave induced density perturbations in the Na layer. This technique can be used with airborne lidars, multiple ground-based lidars, and steerable lidars. Several examples have been presented to show the expected measurement errors for the wave parameters. The technique is applied to the airborne Na lidar data obtained during the westward flight conducted in November 1986. During the flight, strong wave perturbations were observed in the Na layer near the Pacific Coast over a horizontal distance of nearly 700 km. The intrinsic horizontal wavelength of this wave was estimated to be about 85 km, and the vertical wavelength was 4.1 km. The intrinsic period was about 1.7 hours, and the propagation direction was almost due south.

Kinetic energy horizontal and vertical wavenumber spectra of horizontal winds are presented in Chapter 3. The spectra have been inferred from the airborne lidar data collected during the eastward and westward flights. The average slope of the horizontal wavenumber spectra was -1.25 at horizontal scales ranging from 70 to 700 km, and the average slope of the vertical wavenumber spectra was -2.67 at vertical scales from 2 to 10 km. The altitude range of the measurements was approximately 82 to 101 km. The slopes of the horizontal wavenumber spectra computed only for the bottomside (82-90 km) of the Na layer were consistently steeper than those computed for the topside (90-101 km). The average slope of the bottomside horizontal wavenumber spectra was -1.51, and that of the topside spectra was -0.97.

Internal gravity waves appear to be responsible for major features of the airborne Na lidar data. The observed features include the systematic horizontal and vertical variations of the Na density profiles, the horizontal variations of the centroid height, the presence of distinct spectral

peaks in the horizontal wavenumber spectra, and Doppler-shifting of these spectral peaks. It is difficult to interpret these observed features in terms of turbulence. However, it seems possible that the dominant density perturbations observed in the lidar data at horizontal scales in the range from 20 to 2000 km are due to gravity waves.

The slopes of the vertical wavenumber spectra agree well with the slopes calculated from radar observations. The slopes of the horizontal wavenumber spectra of the airborne lidar data are shallower than the slopes of the GASP and shuttle re-entry spectra. In general, the rms horizontal wind velocities measured with the airborne lidar increased with time and with longitude from the Pacific Coast to the Great Plains. These rms horizontal wind velocities were comparable with those measured with ground-based lidars in Hawaii, Illinois, and Maryland. This appears to indicate that the magnitudes of the horizontal winds over the Pacific Ocean are comparable with those over the land masses.

During the eastward flight, two quasi-monochromatic waves were observed. One wave had a much longer zonal wavelength. The parameters of these waves were computed by using the Doppler shifted zonal wavelengths measured during the eastbound and westbound flight legs. The longer wavelength wave had an intrinsic zonal wavelength of 772 km, intrinsic zonal phase velocity of 35 m s^{-1} westward, and intrinsic period of 6.1 hours. This wave appears to be propagating almost due west, and is believed to be an atmospheric tide. The shorter wavelength wave had an intrinsic zonal wavelength of 263 km, zonal phase velocity of 43 m s^{-1} westward, and period of 1.7 hours. This wave also appears to be propagating westward.

In Chapter 4, the results of a joint lidar/radar campaign conducted in November, 1986 are presented. The lidar campaign included the airborne observations and ground-based observations at Broomfield and Denver, Colorado. The radar observations were obtained at Platteville, Colorado with the ST radar. The radar observations showed a dominant variation in the horizontal wind perturbations with a period of 6 hours. The ground-based and airborne lidar observations also showed dominant Na density variations by waves with periods of 6 hours.

These 6-hour period variations observed with the radar, ground-based lidar, and airborne lidar were most dominant at the altitudes which correspond to the bottomside of the layer. It appears that the 6-hour period waves observed with the ground-based and airborne lidars are responsible for the horizontal wind variation with a period of 6 hours observed with the radar. The data obtained with both the ground-based and airborne lidars also exhibited dominant Na density variations by waves with periods of approximately 2 hours. These 2-hour period waves were also dominant only on the bottomside of the layer.

The characteristics of sporadic Na layers observed above Mauna Kea, Hawaii (20°N, 155°W) are described in Chapter 5. These layers were observed on a total of 16 occasions during 30 hours of lidar measurements from January 17 to 22, 1987. The most prominent sporadic layer, which formed on the night of January 21, exhibited a peak density of $2.8 \times 10^4 \text{ cm}^{-3}$ near 96 km with a full width of about 2 km. The rapid growth and decay of the layer lasted for about 40 min. The apparent Na production and decay rates of this layer were approximately $60 \text{ cm}^{-3} \text{ s}^{-1}$. Even after the decay, the narrow layer was observed continuously for almost 8 hours. During this period the layer moved downward steadily with a mean apparent velocity of 31 cm s^{-1} . The most significant characteristic of the 16 sporadic layers appears to be occurrence times. The layers formed either in the late evening between 2100 and 2330 LST or in the early morning between 0300 and 0600 LST. The layers of the early morning began forming within a short time span of 15 min from 0253 to 0308 LST on three different days. The mean time of the maximum peak density of the early morning layers occurred about 6 hours after that of the late evening layers. The mechanisms responsible for creating these layers appear to be related to diurnal tides and sporadic E layers.

From July, 1987 to April, 1988, the UIUC group conducted a total of five Na lidar campaigns at Nordlysstasjonen, Svalbard (78°N, 16°E), Norway. The characteristics of the Na layer observed at Svalbard are described in Chapter 6. The seasonal variations of the abundance measured at Svalbard show an annual oscillation with a broad maximum in January to April

followed by a distinct minimum in June. The column abundance measured in June was $\sim 6 \times 10^8$ cm^{-2} , a value that is almost 5 times lower than typical summertime Na abundances measured at mid-latitudes. The centroid measured at Svalbard shows the semi-annual oscillations with maxima near equinox in March and September and minima near solstice in December and June. This semi-annual oscillations are quite similar to those observed at Urbana. During the winter months, the centroid height observed at Svalbard was generally lower than the centroid observed at Urbana. This could be the result of a stronger downward motion of the atmosphere over the high latitude site of Svalbard during the winter months. The rms width measured at Svalbard shows an annual oscillation with a maximum near February and a minimum near September. The characteristics of sporadic Na layers observed at Svalbard are quite similar to those observed at Hawaii (Chapter 5).

7.2 Recommendations for Future Work

The data analysis technique described in Chapter 2 can be applied to airborne, multiple ground-based, and steerable lidar experiments. In June 1981, steerable lidar experiments were conducted at the Goddard Space Flight Center in Maryland. The new data analysis technique can be applied to these data in order to estimate gravity wave parameters. Additional steerable measurements with either the Candela or the CEDAR lidars are also recommended. Joint observations of a steerable lidar and an MST radar will be quite valuable, because the radar can measure the background winds, and the steerable lidar can measure the propagation directions of gravity waves. Comparison of these data can help determine wave propagation characteristics at the altitudes of the Na layer.

The airborne campaign which was conducted in November 1986 was very successful. The flight paths were selected to investigate longitudinal characteristics of the Na layer and gravity waves. Additional airborne campaigns are recommended to investigate the latitudinal

characteristics of the layer and gravity waves. Triangular flight patterns are also recommended to measure the wave propagation directions.

Dominant waves with periods of approximately 2 hours have been observed often with the Na lidars at Urbana, Illinois (40°N) [Gardner *et al.*, 1986; Gardner and Voelz, 1987; Gardner, 1989], at Broomfield, Colorado (40°N) [Kwon *et al.*, 1989b], during the eastward flight on November 15-16, 1986 [Kwon *et al.*, 1989a], and during the westward flight on November 17-18 [Kwon *et al.*, 1989c]. It appears that these 2-hour period waves often dominate the bottomside (80 - 90 km) of the Na layer at mid-latitudes over North America. Studies of the 2-hour period waves observed at Urbana are recommended to characterize the waves.

The mechanisms responsible for creating and dissipating the sporadic Na layers are not well understood yet. The sporadic Na layers have been observed often at the low latitude site of Hawaii [Kwon *et al.*, 1988] and the high latitude site of Svalbard (Chapter 6), but very rarely at the mid-latitude site of Urbana [Senft *et al.*, 1989]. In November 1987 and January 1988, simultaneous observations of lidar and airglow were conducted at Svalbard. The airglow observations were made at the wavelengths of Na, OH(6-2), and Oxygen atmospheric (0-1) bands. The increases in the intensity of Na and OH airglow were observed almost simultaneously with the appearances of sporadic Na layers [private communication with Roger Smith, University of Alaska]. Another recent observations of sporadic Na layers at Arecibo (18°N, 67°W) in January, 1989 have revealed that the altitudes of sporadic Na layers coincided with the altitudes of sporadic E layers. These observations at Svalbard and Arecibo are quite unique and should help determine the mechanisms responsible for creating and dissipating the sporadic Na layers.

In January and March 1986, several daytime observations were obtained at Urbana in 1986. More daytime observations at Urbana are strongly recommended to characterize the mean diurnal variations of the Na layer and to study the climatology of tidal waves.

APPENDIX I

RMS ERRORS IN GRAVITY WAVE INTRINSIC PARAMETERS

FOR AIRBORNE LIDAR OBSERVATIONS OVER A

CIRCULAR FLIGHT PATH

Consider an aircraft flying over a circular path of radius R during a total observation period of T_{ob} . The ground track of the flight is illustrated in Figure 2.2. Assume that a total of n measurements equally spaced along the flight path is obtained. Let the reference time be $t_0 = T_{ob}/2$, the reference horizontal position be $(x_0, y_0) = (0, 0)$, and the reference altitude be z_0 . The i^{th} measurement referenced to (t_0, x_0, y_0, z_0) is

$$\Delta t_i = t_i - t_0 = \frac{T_{ob}}{n} i - \frac{T_{ob}}{2} \quad (I.1)$$

$$\Delta x_i = x_i - x_0 = R \cos\left(\frac{2\pi i}{n}\right) \quad (I.2)$$

$$\Delta y_i = y_i - y_0 = R \sin\left(\frac{2\pi i}{n}\right) \quad (I.3)$$

$$\Delta z_i = z_i - z_0 \quad (I.4)$$

where t_i , x_i , y_i , and z_i correspond to the time, x-coordinate, y-coordinate, and altitude of the i^{th} measurement, and $1 \leq i \leq n$. These measurements are substituted into Equation (2.3), and the resulting system of equations is written in matrix form.

$$\mathbf{Z} = \mathbf{U} \mathbf{A} \quad (I.5)$$

$$\text{where } \mathbf{Z} = [\Delta z_1 \ \Delta z_2 \ \dots \ \Delta z_n]^T \quad (I.6)$$

$$U = \begin{bmatrix} \Delta t_1 & \Delta x_1 & \Delta y_1 & 1 \\ \Delta t_2 & \Delta x_2 & \Delta y_2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \Delta t_n & \Delta x_n & \Delta y_n & 1 \end{bmatrix} \quad (I.7)$$

$$A = [a_1 \ a_2 \ a_3 \ a_4]^T \quad (I.8)$$

In order to estimate the errors in the gravity wave parameters, the $U^T U$ matrix and the coefficient covariance matrix have to be calculated.

$$U^T U = \begin{bmatrix} \Sigma t_i^2 & \Sigma t_i x_i & \Sigma t_i y_i & \Sigma t_i \\ \Sigma t_i x_i & \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \\ \Sigma t_i y_i & \Sigma x_i y_i & \Sigma y_i^2 & \Sigma y_i \\ \Sigma t_i & \Sigma x_i & \Sigma y_i & n \end{bmatrix} \quad (I.9)$$

where the summations are over $1 \leq i \leq n$. In this case,

$$U^T U \approx \begin{bmatrix} \frac{nT_{ob}^2}{12} & 0 & \frac{-nRT_{ob}}{2\pi} & 0 \\ 0 & \frac{nR^2}{2} & 0 & 0 \\ \frac{-nRT_{ob}}{2\pi} & 0 & \frac{nR^2}{2} & 0 \\ 0 & 0 & 0 & n \end{bmatrix} \quad (I.10)$$

The coefficient covariance matrix is given as $C = [U^T U]^{-1} \text{Var}(z)$,

$$C = \begin{bmatrix} \frac{12\pi^2}{(\pi^2-6)nT_{ob}^2} & 0 & \frac{12\pi}{(\pi^2-6)nRT_{ob}} & 0 \\ 0 & \frac{2}{nR^2} & 0 & 0 \\ \frac{12\pi}{(\pi^2-6)nRT_{ob}} & 0 & \frac{2\pi^2}{(\pi^2-6)nR^2} & 0 \\ 0 & 0 & 0 & \frac{1}{n} \end{bmatrix} \text{Var}(z) \quad (I.11)$$

Then, by using Equations (2.18) and (2.19) the rms errors in the gravity wave parameters are

$$\begin{aligned} \text{Std}(\alpha_w) &= \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \sqrt{\frac{\pi^2}{\pi^2-6} \sin^2 \alpha_w + \cos^2 \alpha_w} \\ &= \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \end{aligned} \quad (I.12)$$

$$\begin{aligned} \frac{\text{Std}\left(\frac{\lambda_h}{\lambda_z}\right)}{\left(\frac{\lambda_h}{\lambda_z}\right)} &= \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \sqrt{\sin^2 \alpha_w + \frac{\pi^2}{\pi^2-6} \cos^2 \alpha_w} \\ &= \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \end{aligned} \quad (I.13)$$

APPENDIX II

RMS ERRORS IN GRAVITY WAVE INTRINSIC PARAMETERS FOR MULTIPLE GROUND-BASED LIDAR OBSERVATIONS

Consider a configuration of three ground-based lidars located at the corners of an equilateral triangle with sides of length R as illustrated in Figure 2.3. The horizontal coordinates of the three lidar sites are

$$\text{Site A: } (x_a, y_a) = \left(0, \frac{R}{\sqrt{3}} \right) \quad (\text{II.1})$$

$$\text{Site B: } (x_b, y_b) = \left(\frac{-R}{2}, \frac{-R}{2\sqrt{3}} \right) \quad (\text{II.2})$$

$$\text{Site C: } (x_c, y_c) = \left(\frac{R}{2}, \frac{-R}{2\sqrt{3}} \right) \quad (\text{II.3})$$

Assume that a total of n measurements are obtained simultaneously at each of the three lidar sites over a total observation period of T_{ob} . Let the reference time be $t_0 = T_{ob}/2$, the reference horizontal position be $(x_0, y_0) = (0, 0)$, and the reference altitude be z_0 . The differences between the observations at site A and the reference data point are

$$\Delta t_{ai} = t_{ai} - t_0 = \frac{T_{ob}}{n} i - \frac{T_{ob}}{2} \quad (\text{II.4})$$

$$\Delta x_{ai} = x_a \quad (\text{II.5})$$

$$\Delta y_{ai} = y_a \quad (\text{II.6})$$

$$\Delta z_{ai} = z_{ai} - z_0 \quad (\text{II.7})$$

where z_{ai} is the i^{th} observation of the altitude of a density maximum or minimum made at time t_{ai} at site A. Data from sites B and C are written using similar notations. These data are substituted into Equation (2.3), and the resulting system of equations is written in matrix form.

$$Z = U A \quad (\text{II.8})$$

$$\text{where } Z = [\Delta z_{a1} \dots \Delta z_{an} \Delta z_{b1} \dots \Delta z_{bn} \Delta z_{c1} \dots \Delta z_{cn}]^T \quad (\text{II.9})$$

$$U = \begin{bmatrix} \Delta t_{a1} & \Delta x_{a1} & \Delta y_{a1} & 1 \\ : & : & : & : \\ \Delta t_{an} & \Delta x_{an} & \Delta y_{an} & 1 \\ \Delta t_{b1} & \Delta x_{b1} & \Delta y_{b1} & 1 \\ : & : & : & : \\ \Delta t_{bn} & \Delta x_{bn} & \Delta y_{bn} & 1 \\ \Delta t_{c1} & \Delta x_{c1} & \Delta y_{c1} & 1 \\ : & : & : & : \\ \Delta t_{cn} & \Delta x_{cn} & \Delta y_{cn} & 1 \end{bmatrix} \quad (\text{II.10})$$

In order to estimate the errors in the gravity wave parameters, the $U^T U$ matrix and the coefficient covariance matrix have to be calculated.

$$U^T U = \begin{bmatrix} \frac{nT_{ob}^2}{4} & 0 & 0 & 0 \\ 0 & \frac{nR^2}{2} & 0 & 0 \\ 0 & 0 & \frac{nR^2}{2} & 0 \\ 0 & 0 & 0 & 3n \end{bmatrix} \quad (\text{II.11})$$

And the coefficient covariance matrix is

$$C = \begin{bmatrix} \frac{4}{nT_{ob}^2} & 0 & 0 & 0 \\ 0 & \frac{2}{nR^2} & 0 & 0 \\ 0 & 0 & \frac{2}{nR^2} & 0 \\ 0 & 0 & 0 & \frac{1}{3n} \end{bmatrix} \text{Var}(z) \quad (\text{II.12})$$

By using Equations (2.18) and (2.19) the rms errors in the gravity wave parameters are given as

$$\text{Std}(\alpha_w) = \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \quad (\text{II.13})$$

$$\frac{\text{Std}\left(\frac{\lambda_h}{\lambda_z}\right)}{\left(\frac{\lambda_h}{\lambda_z}\right)} = \frac{\lambda_h}{\lambda_z} \frac{\text{Std}(z)}{R} \sqrt{\frac{2}{n}} \quad (\text{II.14})$$

APPENDIX III
CHARACTERISTICS OF SPORADIC SODIUM LAYERS OBSERVED
AT MAUNA KEA OBSERVATORY , HAWAII

Sporadic Layer Number	Date	Time Period (LST)	Duration (hour:min)	Beginning and Ending Altitudes (km)	Sporadic E layers
1	Jan 21	2211 - 0230	4:19	95.40, 91.50	Yes
2	Jan 21	2217 - 2223	0:06	97.80, 96.90	Yes
3	Jan 21	2226 - 2235	0:09	94.50, 93.00	Yes
4	Jan 21	2200 - 2213	0:13	98.25, 96.45	Yes
5	Jan 22	0253 - 0600 ^a	3:07 ^b	91.50, 87.30	No
6	Jan 19	2114 ^c - 2117 ^a	0:03 ^b	96.60, 96.60	Yes
7	Jan 20	0256 - 0621 ^a	3:25 ^b	96.45, 93.15	No
8	Jan 21	2112 - 2131	0:19	94.35, 94.20	No
9	Jan 21	2226 - 2326	1:00	98.40, 100.05	Yes
10	Jan 20	2253 - 2354	1:01	96.45, 93.45	No
11	Jan 18	0308 - 0453	1:45	92.55, 93.75	No
12	Jan 21	2251 - 2330 ^a	0:39 ^b	103.35, 102.00	Yes
13	Jan 20	2052 ^c - 2231	1:39 ^b	92.25, 92.55	No
14	Jan 20	0305 - 0356	0:51	94.50, 94.05	No
15	Jan 21	2226 - 2248	0:22	100.80, 101.25	Yes
16	Jan 21	2140 - 2200	0:20	104.40, 102.15	No

^aSporadic layer present when data collection stopped.

^bMeasurement of the duration was impossible because sporadic layers already formed when data collection began or was present when data collection stopped.

^cSporadic layer already formed when data collection began.

Sporadic Layer Number	Strength Factor	Maximum Density at Peak (cm ⁻³)	Time of Maximum Density (LST)	Altitude of Maximum Density (km)	Sporadic Layer Abundance (cm ⁻²)
1	13.8	27890	2224	95.70	6.7x10 ⁹
2	14.1	17680	2218	97.50	2.0x10 ⁹
3	5.1	12800	2226	94.50	1.5x10 ⁹
4	6.2	9640	2211	97.05	1.3x10 ⁹
5	6.8	9280	0407	89.85	1.9x10 ⁹
6	NA	6600	NA	NA	7.0x10 ⁸
7	4.6	5870	0451	93.45	1.9x10 ⁹
8	1.6	4760	2116	94.20	7.2x10 ⁸
9	4.3	4180	2226	98.40	3.7x10 ⁸
10	2.0	4160	2317	95.40	9.6x10 ⁸
11	1.9	3400	0312	92.85	6.5x10 ⁸
12	16.3	3390	2321	101.70	5.0x10 ⁸
13	2.0	3030	2101	92.40	6.9x10 ⁸
14	1.7	2860	0345	94.20	4.2x10 ⁸
15	15.4	2350	2229	100.50	2.4x10 ⁸
16	27.5	1670	2158	102.15	3.2x10 ⁸

NA, not available.

Sporadic Layer Number	Average Na Production Rate (cm ⁻³ s ⁻¹)	Formation Period (min)	Average Na Decay Rate (cm ⁻³ s ⁻¹)	Decay Period (min)	Mean Vertical Velocity of Layer Peak (cm s ⁻¹)
1	28.4	12.8	12.64	27.5	-28.5
2	92.7	1.8	22.3	3.7	-274
3	NA	NA	17.3	9.2	-293
4	8.06	11.0	35.4	1.8	-205
5	1.03	73.3	2.18	60.5	-35.1
6	NA	NA	NA	NA	NA
7	2.71	18.3	0.52	55.0	-33.2
8	2.62	3.7	1.17	14.7	3.30
9	10.9	1.8	1.52	34.8	155
					-35.9
10	2.64	7.3	1.95	16.5	64.9
11	2.03	3.7	0.33	82.5	12.4
12	5.15	9.2	NA	NA	-184
					-22.8
13	NA	NA	0.28	60.5	2.40
14	0.23	40.3	0.46	9.2	-14.7
15	1.75	3.7	1.86	12.8	25.8
16	3.23	3.7	4.52	1.8	-181

Sporadic Layer Number	Mean Full Width at 80% (km)	Mean Full Width Broadening Rate (cm s ⁻¹)	Mean Top-half Width (km)	Mean Top-half Width Broadening Rate (cm s ⁻¹)	Mean Bottom-half Width (km)	Mean Bottom-half Width Broadening Rate (cm s ⁻¹)
1	1.25	8.17	0.74	7.41	0.51	0.76
2	0.68	27.5	0.34	-13.7	0.34	41.1
3	0.45 ^d	18.4	0.27 ^d	25.8	0.18 ^d	-7.37
4	0.88	11.5	0.35	6.57	0.53	4.93
5	1.21	20.1	0.64	11.6	0.57	8.46
6	0.60	NA	0.30	NA	0.30	NA
7	1.29	11.8	0.66	7.20	0.64	4.55
8	0.56 ^d	23.6	0.21 ^d	9.89	0.34 ^d	13.7
9	0.69	4.67	0.32	2.71	0.36	1.95
	0.65	-8.06	0.33	-9.44	0.33	1.39
10	1.54	11.7	0.82	12.1	0.72	-0.34
11	1.22	1.67	0.65	0.80	0.57	0.87
12	0.87	7.35	0.45	18.4	0.42	-11.1
	0.81	2.59	0.43	0.72	0.39	3.31
13	1.26	-0.33	0.71	-1.86	0.55	1.53
14	0.47 ^d	1.19	0.27 ^d	-0.72	0.20 ^d	1.92
15	0.67	3.26	0.31	6.03	0.35	-2.77
16	0.90	-9.45	0.46	-20.8	0.44	11.4

^dWidth determined at 95% of the peak density.

REFERENCES

- Avery, S. K., and D. Tetenbaum, Simultaneous sodium and wind measurements in the upper mesosphere using the Urbana meteor radar and lidar systems, *J. Atmos. Terr. Phys.*, **45**, 753-764, 1983.
- Balsley, B. B., and D. A. Carter, The spectrum of atmospheric velocity fluctuations at 8 km and 86 km, *Geophys. Res. Lett.*, **9**, 465-468, 1982.
- Balsley, B. B., and R. Garello, The kinetic energy density in the troposphere, stratosphere and mesosphere: A preliminary study using the Poker Flat radar in Alaska, *Radio Sci.*, **20**, 1355-1362, 1985.
- Batista, P. P., B. R. Clemesha, D. M. Simonich, and V. W. J. H. Kirchhoff, Tidal oscillations in the atmosphere sodium layer, *J. Geophys. Res.*, **90**, 3881-3888, 1985.
- Bernard, R., J. L. Fellows, M. Massebeuf, and M. Glass, Simultaneous meteor radar observations of Monpazier (France, 44°N) and Punta Borinquen (Puerto Rico, 18°N), I, Latitudinal variations of atmospheric tides, *J. Atmos. Terr. Phys.*, **43**, 525-533, 1981.
- Beatty, T. J., R. E. Bills, K. H. Kwon, and C. S. Gardner, CEDAR lidar observations of sporadic Na layers at Urbana, Illinois, *Geophys. Res. Lett.*, **15**, 1137-1140, 1988.
- Blamont, J. E., and T. M. Donahue, The airglow of the sodium D lines, *J. Geophys. Res.*, **66**, 1407-1423, 1961.
- Bowman, M. R., A. J. Gibson, and M. C. W. Sandford, Atmospheric sodium measured by a tuned laser radar, *Nature*, **221**, 456-457, 1969.
- Burnett, C. R., R. W. Lasher, A. S. Miskin, and V. L. Sides, Spectroscopic measurement of sodium airglow: Absence of a large diurnal variation, *J. Geophys. Res.*, **80**, 1837-1844, 1975.
- Chiu, Y. T., and B. K. Ching, The response of atmospheric and lower ionospheric layer structures to gravity waves, *Geophys. Res. Lett.*, **5**, 539-542, 1978.
- Clemesha, B. R., V. W. J. H. Kirchhoff, D. M. Simonich, and H. Takahashi, Evidence of an extraterrestrial source for the mesospheric sodium layer, *Geophys. Res. Lett.*, **5**, 873-876, 1978.
- Clemesha, B. R., V. W. J. H. Kirchhoff, D. M. Simonich, H. Takahashi, and B. P. Batista, Spaced lidar and nightglow observations of an atmospheric sodium enhancement, *J. Geophys. Res.*, **85**, 3480-3484, 1980.
- Clemesha, B. R., V. W. J. H. Kirchhoff, D. M. Simonich, and P. P. Batista, Mesospheric winds from lidar observations of atmospheric sodium, *J. Geophys. Res.*, **86**, 868-870, 1981.
- Donahue, T. M., and R. R. Meier, Distribution of sodium in the daytime upper atmosphere as measured by a rocket experiment, *J. Geophys. Res.*, **72**, 2803-2829, 1967.

- Fricke, K. H., and U. von Zahn, Mesopause temperature derived from probing the hyperfine structure of the D₂ resonance line of sodium by lidar, *J. Atmos. Terr. Phys.*, 47, 499-512, 1985.
- Fritts, D. C., Gravity wave saturation in the middle atmosphere: A review of theory and observations, *Rev. Geophys.*, 22, 275-308, 1984.
- Fritts, D. C., M. A. Geller, B. B. Balsley, M. L. Chanin, I. Hirota, J. R. Holton, S. Kato, R. S. Lindzen, M. R. Schoeber, R. A. Vincent, and R. F. Woodman, Research status and recommendations from the Alaska workshop on gravity waves and turbulence in the middle atmosphere, *Bull. Am. Meteorol. Soc.*, 65, 149-159, 1984.
- Fritts, D. C., R. C. Blanchard, and L. Coy, Gravity wave structure between 60 and 90 km inferred from space shuttle re-entry data, *J. Atmos. Sci.*, in press, 1989.
- Gadsen, M., and C. M. Purdy, Observations of the sodium dayglow, *Ann. Geophys.*, 26, 43-51, 1970.
- Gage, K. S., and G. D. Nastrom, On the spectrum of atmosphere velocity fluctuations seen by MST/ST radar and their interpretation, *Radio Sci.*, 20, 1339-1347, 1985.
- Gage, K. S., and G. D. Nastrom, Theoretical interpretation of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft during GASP, *J. Atmos. Sci.*, 43, 729-740, 1986.
- Gardner, C. S., C. F. Sechrist, Jr., and J. L. Bufton, Steerable lidar studies of the mesospheric sodium layer structure, *Proc. 11th Int. Laser Radar Conf.*, Madison, WI, 1982.
- Gardner, C. S., and J. D. Shelton, Density response of neutral atmospheric layers to gravity wave perturbations, *J. Geophys. Res.*, 90, 1745-1754, 1985.
- Gardner, C. S., D. G. Voelz, C. F. Sechrist, Jr., and A. C. Segal, Lidar studies of the nighttime sodium layer over Urbana, Illinois 1. Seasonal and nocturnal variations, *J. Geophys. Res.*, 91, 13659-13673, 1986.
- Gardner, C. S., and D. G. Voelz, Lidar studies of the nighttime sodium layer over Urbana, Illinois. 2. Gravity waves and tides, *J. Geophys. Res.*, 92, 4673-4694, 1987.
- Gardner, C. S., D. C. Senft, K. H. Kwon, R. E. Bills, and K. Henrikson, Lidar observations of sodium depletion in the summertime arctic mesosphere, AGU fall meeting, San Francisco, Calif., Dec. 7-11, 1987.
- Gardner, C. S., D. C. Senft, and K. H. Kwon, Lidar observations of substantial sodium depletion in the summertime arctic mesosphere, *Nature*, 332, 142-144, 1988.
- Gardner, C. S., Sodium resonance fluorescence lidar applications in atmospheric science and astronomy, *Proc. IEEE*, in press, 1989.
- Gardner, C. S., M. S. Miller, and C. H. Liu, Rayleigh lidar observations of gravity wave activity in upper stratosphere at Urbana, Illinois, *J. Atmos. Sci.*, in press, 1989.

- Gardner, C. S., and D. C. Senft, Sodium lidar studies of gravity wave winds and spectra in the mesosphere, in preparation, April, 1989.
- Gibson, A. J., and M. C. W. Sandford, The seasonal variations of the nighttime sodium layer, *J. Atmos. Terr. Phys.*, **33**, 1675-1684, 1971.
- Gibson, A. J., L. Thomas, and S. K. Bhattachacharyya, Laser observations of the ground-state hyperfine structure of sodium and of temperatures in the upper atmosphere, *Nature*, **281**, 131-132, 1979.
- Hines, C. O., Internal atmospheric gravity waves at ionospheric heights, *Can. J. Phys.*, **38**, 1441-1481, 1960.
- Holton, J. R., The role of gravity wave induced drag and diffusion in the momentum budget of the mesosphere, *J. Atmos. Sci.*, **39**, 791-799, 1982.
- Holton, J. R., The influence of gravity wave breaking on the general circulation of the middle atmosphere, *J. Atmos. Sci.*, **40**, 2497-2507, 1983.
- Hunten, D. M., and L. Wallace, Rocket measurements of the sodium dayglow, *J. Geophys. Res.*, **72**, 69-79, 1967.
- Jegou, J. P., C. Granier, M. L. Chanin, and G. Megie, General theory of the alkali metals present in the Earth's upper atmosphere, I, Flux model: Chemical and dynamical processes, *Ann. Geophys.*, **3**, 163-176, 1985a.
- Jegou, J. P., C. Granier, M. L. Chanin, and G. Megie, General theory of the alkali metals present in the Earth's upper atmosphere, II, Seasonal and meridional variations, *Ann. Geophys.*, **3**, 299-312, 1985b.
- Juramy, P., M. L. Chanin, G. Megie, G. F. Toulinov, and Y. P. Doudoladov, Lidar sounding of the mesospheric sodium layer at high latitudes, *J. Atmos. Terr. Phys.*, **43**, 209-215, 1981.
- Kirchhoff, V. W. J. H., Theory of the atmospheric sodium layer: a review, *Can. J. Phys.*, **64**, 1664-1672, 1986.
- Kwon, K. H., C. S. Gardner, D. C. Senft, F. L. Roesler, and J. Harlander, Daytime lidar measurements of tidal winds in the mesospheric sodium layer at Urbana, Illinois, *J. Geophys. Res.*, **92**, 8781-8786, 1987.
- Kwon, K. H., D. C. Senft, and C. S. Gardner, Lidar observations of sporadic sodium layers at Mauna Kea Observatory, Hawaii, *J. Geophys. Res.*, **93**, 14199-14208, 1988.
- Kwon, K. H., D. C. Senft, and C. S. Gardner, Airborne sodium lidar observations of horizontal and vertical wavenumber spectra of mesopause density and wind perturbations, submitted to *J. Geophys. Res.*, Feb. 1989a.
- Kwon, K. H., C. S. Gardner, S. K. Avery, J. P. Avery, and C. A. Whitmord, Correlative radar and airborne sodium lidar observations of the vertical and horizontal structure of gravity waves and tides near the mesopause, in preparation, April, 1989b.

- Kwon, K. H., and C. S. Gardner, Airborne sodium lidar measurements of gravity wave intrinsic parameters, in preparation, April, 1989c.
- Lindzen, R. S., Turbulence and stress owing to gravity wave and tidal breakdown, *J. Geophys. Res.*, **86**, 9707-9714, 1981.
- Mathews, J. D., Measurements of the diurnal tides in the 80-100 km altitude range at Arecibo, *J. Geophys. Res.*, **81**, 4671-4677, 1976.
- Manson, A.H. and C.E. Meek, Gravity wave propagation characteristics (60-120 km) as determined by the Saskatoon MF radar (Gravnet) system: 1983-85 at 52°N, 107°W, *J. Atmos. Sci.*, **45**, 932-946, 1988.
- Meek, C. E., I. M. Reid, and A. H. Manson, Observations of mesospheric wind velocities 1. Gravity wave horizontal scales and phase velocities determined from spaced wind observations, *Radio Sci.*, **20**, 1363-1382, 1985.
- Megie, G., and J. E. Blamont, Laser sounding of atmospheric sodium: Interpretation in terms of global atmospheric parameters, *Planet. Space Sci.*, **25**, 1093-1109, 1977.
- Megie G., M. L. Chanin, G. F. Tulinov, Y. P. Dudoladov, High latitude measurements of the atomic sodium concentration and neutral temperature at the mesopause level by lidar technique, *Planet. Space Sci.*, **26**, 509-511, 1978.
- Neuber, R., P. von der Gathen, and U. von Zahn, Altitude and temperature of the mesopause at 69°N latitude in winter, *J. Geophys. Res.*, **93**, 11093-11101, 1988.
- Nomura, A., T. Kano, Y. Iwasaka, H. Fukunishi, T. Hirasawa, and S. Kawaguchi, Lidar observations of the mesospheric sodium layer at Syowa Station, Antarctica, *Geophys. Res. Lett.*, **14**, 700-703, 1987.
- Philbrick, C. R., J. Barnett, R. Gerndt, D. Offermann, W. R. Pendleton, Jr., P. Schlyter, J. F. Schmidlin, and G. Witt, Temperature measurements during the CAMP program, *Adv. Space Res.*, **4**, 153-156, 1984.
- Reid, I.M., and R.A. Vincent, Measurements of the horizontal scales and phase velocities of short period mesospheric gravity waves at Adelaide, Australia, *J. Atmos. Terr. Phys.*, **49**, 1033-1048, 1987.
- Richter, E. S., and C. F. Sechrist, Jr., A cluster ion chemistry for the mesospheric sodium layer, *J. Atmos. Terr. Phys.*, **41**, 579-586, 1979.
- Rowlett, J. R., C. S. Gardner, E. S. Richter, and C. F. Sechrist, Jr., Lidar observations of wave-like structure in the atmospheric sodium layer, *Geophys. Res. Lett.*, **5**, 683-686, 1978.
- Segal, A. C., D. G. Voelz, C. S. Gardner, and C. F. Sechrist, Jr., Airborne lidar observations of the mesospheric sodium layer, *12th International Laser Radar Conference*, 243-246, G. Megie, Ed., Aix en Provence, France, 1984.
- Senft, D. C., R. L. Collins, and C. S. Gardner, First lidar observations of large sporadic sodium layers at mid-latitudes, submitted to *Geophys. Res. Lett.*, 1989.

- Simonich, D. M., B. R. Clemesha, and V. W. J. H. Kirchhoff, The mesospheric sodium layer at 23°S: Nocturnal and seasonal variations, *J. Geophys. Res.*, **84**, 1543-1550, 1979.
- Slipher, V. M., Emissions in the spectrum of the light of the night sky, *Publs. Astr. Soc. Pacif.*, **41**, 262, 1929.
- Smith, S. A., and D. C. Fritts, Estimation of gravity wave motions, momentum fluxes and induced mean flow accelerations in the winter mesosphere over Poker Flat, Alaska, *21st Conference on Radar Meteorology*, AMS, Edmonton, Canada, 1983.
- Smith, S. A., D. C. Fritts, and T. E. Van Zandt, Comparison of mesosphere wind spectra with a gravity wave model, *Radio Sci.*, **20**, 1331-1338, 1985.
- Swider, W., Enhanced seasonal variations for chemical rates with inverse temperature dependences: Application to seasonal abundance of mesospheric sodium, *Geophys. Res. Lett.*, **12**, 589-591, 1985.
- Thomas, I., A. J. Gibson, and S. K. Bhattacharyya, Lidar observations of a horizontal variation in the atmospheric sodium layer, *J. Atmos. Terr. Phys.*, **39**, 1405-1409, 1977.
- Thomas, L., M. C. Isherwood, and M. R. Bowman, A theoretical study of the height distribution of sodium in the mesosphere, *J. Atmos. Terr. Phys.*, **45**, 587-594, 1983.
- Thompson, L. A., and C. S. Gardner, Experiments on laser guide stars at Mauna Kea Observatory for adaptive imaging in astronomy, *Nature*, **328**, 229-231, 1987.
- Tilgner, C., and U. von Zahn, Average properties of the sodium density distribution as observed at 69°N latitude in winter, *J. Geophys. Res.*, **93**, 8439-8454, 1988.
- Tsuda, T., T. Inoue, D. C. Fritts, T. E. Van Zandt, S. Kato, T. Sato, and S. Fukao, MST radar observations of a saturated gravity wave spectrum, in preparation, 1988.
- Vincent, R. A., and I. M. Reid, HF doppler measurements of mesospheric gravity wave momentum fluxes, *J. Atmos. Sci.*, **40**, 1321-1333, 1983.
- Vincent, R. A., Gravity-wave motions in the mesosphere, *J. Atmos. Terr. Phys.*, **46**, 119-128, 1984.
- Vincent, R. A., and D. C. Fritts, A climatology of gravity wave motions in the mesopause region at Adelaide, Australia, *J. Atmos. Sci.*, **44**, 748-760, 1987.
- von Zahn, U., and R. Neuber, Thermal structure of the high latitude mesopause region in winter, *Beitr. Phys. Atmos.*, **60**, 294-304, 1987.
- von Zahn, U., P. von der Gathen, and G. Hansen, Forced release of sodium from upper atmospheric dust particles, *Geophys. Res. Lett.*, **14**, 76-79, 1987.
- von Zahn, U., and C. Tilgner, The sodium layer at 69°N latitude in wintertime, *Proc. 8th Eur. Space As. Symp. Eur. Rocket and Balloon Prog.*, ESA SP-270, Sunne, Sweden, 17-23, May, 1987.

von Zahn, U., and T. L. Hansen, Sudden neutral sodium layers: A strong link to sporadic E layers, *J. Atmos. Terr. Phys.*, 50, 93-104, 1988.

von Zahn, U., G. Hansen, and H. Kurzawa, Observations of the sodium layer at high latitudes in summer, *Nature*, 331, 594-596, 1988.

Wang, S. T., D. Tetenbaum, B. B. Balsley, R. L. Obert, S. K. Avery, and J. P. Avery, A meteor echo detection and collection system for use on VHF radars, *Radio Sci.*, 23, 46-54, 1988.

VITA

Kang Hyon Kwon was born in Seoul, South Korea on September 25, 1960. He attended the University of Illinois at Urbana-Champaign, from which he received a Bachelor of Science degree in Electrical Engineering in 1984 and a Master of Science degree in Electrical Engineering in 1986.

He was a research assistant at the University from 1984 to 1987, and since 1987 he has been a United States Air Force Laboratory Graduate Fellow. Mr. Kwon is a member of Phi Kappa Phi, Golden Key National Honor Society, and the Institute of Electrical and Electronics Engineers. Mr. Kwon has coauthored the following papers.

1. Kwon, K. H., C. S. Gardner, D. C. Senft, F. L. Roesler, and J. Harlander, Daytime lidar measurements of tidal winds in the mesospheric sodium layer at Urbana, Illinois, *J. Geophys. Res.*, 92, 8781-8786, 1987.
2. Gardner, C. S., D. C. Senft, and K. H. Kwon, Lidar observations of substantial sodium depletion in the summertime arctic mesosphere, *Nature*, 332, 142-144, 1988.
3. Beatty, T. J., R. E. Bills, K. H. Kwon, and C. S. Gardner, CEDAR lidar observations of sporadic Na layers at Urbana, Illinois, *Geophys. Res. Lett.*, 15, 1137-1140, 1988.
4. Kwon, K. H., D. C. Senft, and C. S. Gardner, Lidar observations of sporadic sodium layers at Mauna Kea Observatory, Hawaii, *J. Geophys. Res.*, 93, 14199-14208, 1988.
5. Kwon, K. H., D. C. Senft, and C. S. Gardner, Airborne sodium lidar observations of horizontal and vertical wavenumber spectra of mesopause density and wind perturbations, submitted to *J. Geophys. Res.*, Feb. 1989.
6. Kwon, K. H., C. S. Gardner, S. K. Avery, J. P. Avery, and C. A. Whitmord, Correlative radar and airborne sodium lidar observations of the vertical and horizontal structure of gravity waves and tides near the mesopause, in preparation, April, 1989.
7. Kwon, K. H., and C. S. Gardner, Airborne sodium lidar measurements of gravity wave intrinsic parameters, in preparation, April, 1989.

LATTICE STRUCTURES IN THE IMAGE ALGEBRA
AND APPLICATIONS TO IMAGE PROCESSING

By

JENNIFER L. DAVIDSON

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1989

August 11, 1989

Jennifer L. Davidson
201 Walker Hall
University of Florida
Department of Mathematics
Gainesville, FL 32611


Ms. Judy Conover
Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, OH 45432

Dear Ms. Conover:


Enclosed you will find a copy of my dissertation as requested. I graduate on August 12, 1989, and will finish my time as a graduate student at the University of Florida at that date.

I would like to extend my appreciation and gratitude to UES for administering my grant and fellowships in a very efficient and pleasing manner. All UES employees were very courteous and prompt in responding to my queries.

Sincerely,



Jennifer L. Davidson



THOMAS E. WALSH
DIRECTOR OF RESEARCH

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Gerhard X. Ritter, for teaching me so much about doing research, for allowing me the opportunity to perform independent research on his contract, and for giving me the chance to work in an exciting area of applied mathematics. Without his constant encouragement, I would not have seen the beauty of mathematics, nor would I have succeeded in mathematics the way I did. I thank Dr. David C. Wilson for providing the opportunity of working with him during a summer, and all the help he has given since then. To Dr. Joseph Wilson I extend my deepest gratitude for helping me with questions in computer science. I am also indebted to all the members of my committee for the help and encouragement that they have given me. To my parents go a debt that I can only repay in love: providing me the opportunity to attend a small, private and very good school for my undergraduate education, which was a critical turning point in my life. I would also like to thank Dr. Sam Lambert and Mr. Neal Urquhart of the Air Force Armament Laboratory and Dr. Jasper Lupo of DARPA for partial support of this research under Contract F08635-84-C-0295.

Finally, I acknowledge my debt to the American taxpayers who provided the financial support for the U.S. Fellowship programs, loans, research assistantships and state teaching assistantships which supported me through most of my time in graduate school. I hope to contribute to society so that this support will be justified.

Jennifer L. Davidson

by

Copyright 1989

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF SYMBOLS	vi
ABSTRACT	viii
INTRODUCTION	1
Background of Image Algebra	1
Parallel Image Processing	3
Summary of Results	5
PART I. LATTICE STRUCTURES IN IMAGE ALGEBRA AND OPERATIONS RESEARCH	7
CHAPTERS	
1. THE TWO ALGEBRAS	9
1.1. Image Algebra: Basic Definitions and Notation	9
1.2. Minimax Algebra	24
2. THE ISOMORPHISM	36
PART II. MINIMAX APPLICATIONS TO IMAGE ALGEBRA AND IMAGE PROCESSING	45
3. MAPPING OF MINIMAX ALGEBRA PROPERTIES TO IMAGE ALGEBRA PROPERTIES	46
3.1. Basic Definitions and Properties	46
3.2. Systems of Equations	57
3.3. Rank of Templates	69
3.4. The Eigenproblem in the Image Algebra	79
4. GENERALIZATION OF MATHEMATICAL MORPHOLOGY	84
5. TRANSFORM DECOMPOSITION	100
5.1. New Matrix Decomposition Results	100
5.2. Decomposition of Templates	121
5.3. Applications to Rectangular Templates	127

6. THE DIVISION ALGORITHM	136
6.1. A Division Algorithm in a Non-Euclidean Domain . .	136
6.2. An Image Algebra Division Algorithm	141
7. TWO EXAMPLES	144
7.1. An Operations Research Problem Stated in Image Algebra Notation	144
7.2. An Image Complexity Measure	148
CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH	153
REFERENCES	157
BIOGRAPHICAL SKETCH	161

LIST OF SYMBOLS

Symbol	Explanation
\mathbb{Z}	the set of integers
\mathbb{R}	the set of real numbers
\mathbb{R}^+	the set of non-negative real numbers
F	an arbitrary value set
ϕ	the identity element of F under its group operation
F^n	the Cartesian product of F
2^S	the power set of S (set of all subsets of S)
\emptyset	the empty set
\in, \notin, \subset	is an element of, is not an element of, is a subset of
\cup, \cap	set union, set intersection
$f : X \rightarrow Y$	f is a function from X to Y
f^{-1}	the inverse of function f
$F_{-\infty}$	the set $F \cup \{-\infty\}$
$F_{\pm\infty}$	the set $F \cup \{-\infty, +\infty\}$
$F_{+\infty}$	the set $F \cup \{+\infty\}$
\vee, \wedge	maximum, minimum
$X \setminus Y$	the set difference of X and Y
W, X, Y	coordinate sets
w, x, y	pixel locations
F^X	the set of all functions from X to F
a, b, c	images
$1 \in F^X$	a constant image on X with values at each coordinate 1
$0 \in F^X$	a constant image on X with values at each coordinate 0

Symbol	Explanation
$1 \in (F_{\pm\infty}^X)^X$	a one-point template from X to X with $l_y(x) = \begin{cases} \phi & \text{if } x = y \\ -\infty & \text{otherwise} \end{cases}$
$\Phi \in (F_{\pm\infty}^X)^Y$	the null template with $\Phi_y(x) = -\infty$ for all $y \in Y, x \in X$
$\chi_S(a)$	the characteristic function over set S of image a
$f(a)$	the function f induced pointwise over image a
$S(t_y)$	the support of template $t \in (R^X)^Y$
$S_{-\infty}(t_y)$	the infinite support of template $t \in (F_{\pm\infty}^X)^Y$
$S_{+\infty}(t_y)$	the positive infinite support of template $t \in (F_{\pm\infty}^X)^Y$
t_y	the image function of template t at location y
r, s, t	templates
$(F^X)^Y$	the set of all F valued templates from Y to X
\oplus	generalized convolution
\otimes, \odot	multiplicative maximum, multiplicative minimum
\boxplus, \boxminus	additive maximum, additive minimum
$ S $	the cardinality function, counting the number of elements in set S
$\sum a$	the sum of all pixel values of the image a
$\vee a$	the maximum pixel value in image a
a^*	the additive dual image of the image $a \in R_{\pm\infty}^X$
\bar{a}	the multiplicative dual image of the image $a \in (R_{\pm\infty}^+)^X$
t^*	the additive dual template of the template $t \in (R_{\pm\infty}^X)^Y$
\bar{t}	the multiplicative dual template of the template $t \in ((R_{\pm\infty}^+)^X)^Y$
$t \in M_{mn}$	an $m \times n$ matrix
t^i	the i -th column of the matrix t
t_i	the i -th row of the matrix t
t'	the transpose of the matrix t , or the transpose of the template t
$F_{\pm\infty}$	a blog with group F
$S_{-\infty}(t_i)$	the infinite support of matrix $t \in M_{mn}$ at row i
$S_{+\infty}(t_i)$	the infinite positive support of matrix $t \in M_{mn}$ at row i
$\chi_S^\infty(a)$	the extended characteristic function
\Longleftrightarrow	if and only if

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

LATTICE STRUCTURES IN THE IMAGE ALGEBRA
AND APPLICATIONS TO IMAGE PROCESSING

By

Jennifer L. Davidson

August 1989

Chairman: Dr. Gerhard X. Ritter
Major Department: Mathematics

The research for this dissertation is concerned with the investigation of an algebraic structure, known as image algebra, which is used for expressing algorithms in image processing. The major result of this research is the establishment of a rigorous and coherent mathematical foundation of the subalgebra of the image algebra involving non-linear image transformations. In particular, a classification in the image algebra of a set of non-linear image transformations called lattice transforms is presented, using minimax matrix algebra as a tool. Several applications to image processing problems are discussed. Specifically, in addition to describing several non-linear transform decomposition techniques, the subalgebra is used as a model and a tool for the development of methods to compute lattice transforms locally.

The basic operands and operations of the image algebra and minimax algebra are defined, as well as the relationships between the two algebras. Properties of the minimax algebra including the lattice eigenvalue problem are mapped to the image algebra. Mathematical morphology is shown to be embedded in the image algebra as a special subclass of lattice transforms. Networks of processors are modeled as graphs, and images are represented as functions defined on the nodes of the graph. It is shown that every lattice image-to-image transform can be weakly factored into a product of lattice transformations each of which are implementable on the network if and only if the graph corresponding to the network is strongly connected. Necessary and sufficient conditions are given to decompose a rectangular template into two strip templates. A division algorithm is given which is a generalization of a boolean skeletonizing technique. The transportation problem from linear programming is expressed in the image algebra. A method to produce an image complexity measure is discussed. Most results are given both in image algebra and matrix algebra notation.

INTRODUCTION

Background of the Image Algebra

The results presented in this dissertation reflect the ongoing investigation of the structure of the Air Force image algebra, an algebraic structure specifically designed for use in image processing. The idea of establishing a unifying theory for concepts and operations encountered in image and signal processing has been pursued for a number of years now. It was the 1950's work of von Neumann that inspired Unger to propose a "cellular array" machine on which to implement, in parallel, many algorithms for image processing and analysis [1,2]. Among the machines embodying the original automaton envisioned by von Neumann are NASA's massively parallel processor or MPP [3], and the CLIP series of computers developed by M.J.B. Duff and his colleagues [4,5]. A more general class of cellular array computers are pyramids [6] and the Connection Machine, by Thinking Machines Corporation [7].

Many of the operations that cellular array machines perform can be expressed by a set of primitives, or simple elementary operations. One opinion of researchers who design parallel image processing architectures is that a wide class of image transformations can be represented by a small set of basic operations that induce these architectures. G. Matheron and J. Serra developed a set of two primitives that formed the basis for the initial development of a theoretical formalism capable of expressing a large number of algorithms for image processing and analysis. Special purpose parallel architectures were then designed to implement these ideas. Several systems in use today are Matheron and Serra's Texture

Analyzer [8], the Cytocomputer at the Environmental Research Institute of Michigan (ERIM) [9,10], and Martin Marietta's GAPP [11].

The basic mathematical formalism associated with the above cellular architectures are the concepts of pixel neighborhood arithmetic and mathematical morphology. Mathematical morphology is a mathematical structure used in image processing to express image processing transformations by the use of *structuring elements*, which are related to the shape of the objects to be analyzed. The origins of mathematical morphology lie in work done by H. Minkowski and H. Hadwiger on geometric measure theory and integral geometry [12,13,14]. It was Matheron and Serra who used a few of Minkowski's operations as a basis for describing morphological image transformations [15,16], and then implemented their ideas by building the Texture Analyzer System. Some recent research papers on morphological image processing are Crimmins and Brown [17], Haralick, Lee and Shapiro [18], Haralick, Sternberg and Zhuang [19], and Maragos and Schafer [20,21,22].

It was Serra and Sternberg who first unified morphological concepts into an algebraic theory specifically focusing on image processing and image analysis. The first to use the term "Image Algebra" was, in fact, Sternberg [23,24]. Recently, a new theory encompassing a large class of linear and nonlinear systems was put forth by P. Maragos [25]. However, despite these profound accomplishments, morphological methods have some well known limitations. They cannot, with the exception of a few simple cases, express some fairly common image processing techniques such as Fourier-like transformations, feature extraction based on convolution, histogram equalization transforms, chain-coding, and image rotation. At Perkin-Elmer, P. Miller demonstrated that a straightforward and uncomplicated target detection algorithm, furnished by the U.S. Government, could not be expressed using a morphologically based image algebra [26].

The morphological image algebra is built on the Minkowski addition and subtraction of sets [14], and it is this set-theoretic formulation of its basic operations which does not enable mathematical morphology to be used as a basis for a general purpose algebraic based language for digital image processing. These operations ignore the linear domain, transformations between different domains (spaces of different dimensionalities) and transformations between different value sets, e.g. sets consisting of real, complex, or vector valued numbers. The image algebra which was developed at the University of Florida includes these concepts and also incorporates and extends the morphological operations.

Parallel Image Processing

The processing of images on digital computers requires enormous amounts of time and memory. With the advent of Very Large Scale Integrated (VLSI) circuits, the cellular array of von Neumann became a reality. There are many types of parallel architectures in existence [27], and various ways of categorizing them have been attempted [28]. The general scheme of one popular type of parallel processor is to have many processing elements, or small processors with limited memory, interconnected by communication links. Each processing element can communicate directly with a single controller as well as with a very small number of its neighbors, usually 1 to 8. When the controller gives a signal, all processing elements simultaneously perform some arithmetic and/or logic operation using the values of its neighbors to which it is connected. This type of parallel processor is called a *neighborhood array processor*, as communication links connect the center processor to a small subset of its spatially nearest neighbors. Two typical neighborhood configurations for local interconnection links are given below. The box with the *x* represents the center processor, and the four (or eight) boxes immediately adjacent to *x* represent the four (or eight) processors

with whom x can send and receive information via the communication links. The set of pixel locations relative to the center pixel location, x , form the *local neighborhood* of x .



Figure 1. Two Neighborhood Configurations.

(a) The von Neumann Configuration; (b) The Moore Configuration.

Some of the parallel processors that have been built to implement this type of connection scheme are the MPP, the Distributed Array Processor (ICL DAP) [27,29], the Geometric Arithmetic Parallel Processor (GAPP), and the CLIP4. There are other types of parallel architectures, such as pipeline computers [30] and systolic arrays [31], which differ in construction and implementation of the neighborhood functions. However, the key feature in most of these architectures is that they have a large number of processing elements, each of which communicates directly with only a small subset of the others.

If every value of a transformed image at location x involves arithmetically or logically manipulating information only from pixel locations in the local neighborhood of x , then the transform is called a *local* transform. Assuming that a transform can be described in a local manner, the amount of time to perform a local operation globally on neighborhood array processors is the amount of time it takes one processor to perform it, often a single clock cycle. Certain image transforms which were previously too computationally intensive can now be implemented on parallel and distributed processors.

In general, image transforms are not local, that is, the calculation of a transformed value may depend on input values which are spatially very distant from the processing element. In order to use parallel processors, the transform must first be decomposed into a product of local transforms. The existence of local decompositions is of theoretical and practical interest, and as such provides the main thrust behind the research in this dissertation.

While such parallel architectures are attractive for use in image processing, much research still needs to be done and implementation techniques developed in order to use the architectures most efficiently.

Summary of Results

The results in this dissertation stem from an investigation into the image algebra operations of \boxtimes , \odot , and \vee . A brief description of the image algebra and its use as a model for image processing is presented. A full discussion of the entire image algebra is presented by Ritter et al. [32]. The results given here focus mainly on two non-linear image transform operations whose underlying values have the structure of a lattice. In particular, it is shown that a previously determined, well-defined mathematical structure called the *minimax algebra* can be used to place the study of a wide class of non-linear, lattice-based image transforms on a solid mathematical foundation. We also discuss the mapping of these transforms to certain types of parallel architectures.

It has been well established that the image algebra is capable of expressing all linear transformations [33]. The embedding of linear algebra into the image algebra makes this possible. The major contributions of this thesis are the development of two isomorphisms between the minimax algebra and image algebra which refines the lattice subalgebra of the

image algebra, and the development of new and useful mathematical tools which are of practical use in the area of image processing.

The dissertation is divided into two parts. Part I gives an introduction to the two algebras, the image algebra and minimax algebra. Part II is devoted to presenting new matrix theoretical results which have applications to solving image processing problems.

Specifically, Chapter 1 is of an introductory nature, presenting a historical background of the image algebra and a brief discussion of where lattice structures appear to be useful in mathematically characterizing problems in image processing and operations research.

Chapter 1 also presents a brief introduction to the image algebra as well as to the minimax algebra. We mention that although vector lattices are contained within the image algebra, they have been investigated [34] and will not be discussed here. The isomorphisms which embed the minimax algebra into the image algebra are given in Chapter 2, and mapping of the minimax algebra properties in image algebra notation are presented in Chapter 3. In

Chapter 4 we give the relationship of mathematical morphology to image algebra. In

Chapter 5 we present new matrix theoretical results, which have applications to template decomposition. An algorithm similar to the division algorithm for integers is given both in minimax algebra and image algebra notation in Chapter 6. In Chapter 7 we present the formulation of an operations research in image algebra notation, and give an image complexity algorithm. We then present the conclusions and give suggestions for future research after Chapter 7.

PART I LATTICE STRUCTURES IN IMAGE ALGEBRA AND OPERATIONS RESEARCH

The algebraic structures of early image processing languages such as mathematical morphology had no obvious connection with a lattice structure. Those algebras were developed to express binary image manipulation. As the extension to gray valued images developed, the notions of performing maximums and minimums over a set of numbers emerged. Formal links to lattice structures were not developed until very recently [34], including this dissertation. We present a little background in this area, showing how the lattice properties were inherent in the structures being developed.

The algebraic operations developed by Serra and Sternberg are equivalent and based on the operations of Minkowski addition and Minkowski subtraction of sets in \mathbb{R}^n . Given $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^n$, Minkowski addition is defined by

$$A + B = \{ a + b : a \in A, b \in B \}$$

and Minkowski subtraction is defined by

$$A / B = \overline{A + B},$$

where the bar denotes set complementation. Mathematical morphology was initially developed for boolean image processing, that is, for processing images that have only two values, say 0 and 1. It was eventually extended to include gray-level image processing, that is, images that take on more than two values. The value set underlying the gray value mathematical morphology structure was the set $\mathbb{R}_{-\infty} \equiv \mathbb{R} \cup \{-\infty\}$, the real numbers with $-\infty$ adjoined. Sternberg's functional notation is most often used to express the two morphological operations, as it is simply stated and easy to implement in computer code. The gray value operations of *dilation* and *erosion*, corresponding to Minkowski addition and

subtraction, respectively, are

$$D(x,y) = \max_{i,j} \{A(x-i, y-i) + B(i,j)\}$$

$$E(x,y) = \min_{i,j} [A(x-i, y-i) - B(-i,-j)]$$

respectively, where A and B are real valued functions on \mathbb{R}^2 .

As will be shown, mathematical morphology, which uses the lattice $\mathbb{R}_{-\infty}$, is actually a very special subalgebra of the full image algebra. It is well known that $\mathbb{R}_{\pm\infty} \equiv \mathbb{R} \cup \{+\infty, -\infty\}$ is a complete lattice [36]. The lattice structure provides the basis for categorizing certain classes of image processing problems, which is the main subject of this dissertation.

Operations research has long been known for its class of problems in optimization. A certain type of non-linear operations research problems has been the focus of Cuninghame-Green during his research [37,38]. The types of optimization problems that were considered by this author used arithmetic operations different from the usual multiplication and summation. Some machine scheduling and shortest path problems, for example, could be best characterized by a non-linear system utilizing additions and maximums. A monograph entitled *Minimax Algebra* [39] describes a matrix calculus which uses a special case of what is called a *generalized matrix product* [40], where matrices and vectors assume values from a lattice. A few more conditions such as a group operation on the lattice, and the self-duality of the resulting structure, allow Cuninghame-Green to develop a solid mathematical foundation in which to pose a wide variety of operations research questions. It is an interesting and natural link between matrices with values in a lattice and templates in the image algebra which provides the foundation of this dissertation.

CHAPTER 1 THE TWO ALGEBRAS

1.1. Image Algebra: Basic Definitions and Notation

This section provides the basic definitions and notation that will be used for the image algebra throughout the dissertation. We will define only those image algebra concepts necessary to describe ideas in this document. For a full discourse on all image algebra operands and operations, we refer the reader to a recent publication [31].

The image algebra is a *heterogeneous algebra*, in the sense of Birkhoff [40], and is capable of describing image manipulations involving not only single valued images, but multivalued images. In fact, it has been formally proven that the set of operations is sufficient for expressing any image-to-image transformation defined in terms of a finite algorithmic procedure, and also that the set of operations is sufficient for expressing any image-to-image transformation for an image which has a finite number of gray values [41,42]. We limit our discussion to single valued images in this document, and refer the reader to other publications on multi-valued images [31].

We will present the six basic operands, some of the finitary operations defined between the operands, and also give a few examples.

1.1.1. The Operands of the Image Algebra

The six basic operands are *coordinate sets*, *elements of coordinate sets*, *value sets*, *elements of value sets*, *images*, and *generalized templates*. They are defined as follows.

1. A *coordinate set* X is a subset of \mathbb{R}^k for some k . Two familiar coordinate sets, the *rectangular* and *toroidal* coordinate sets, are shown in Figure 2.

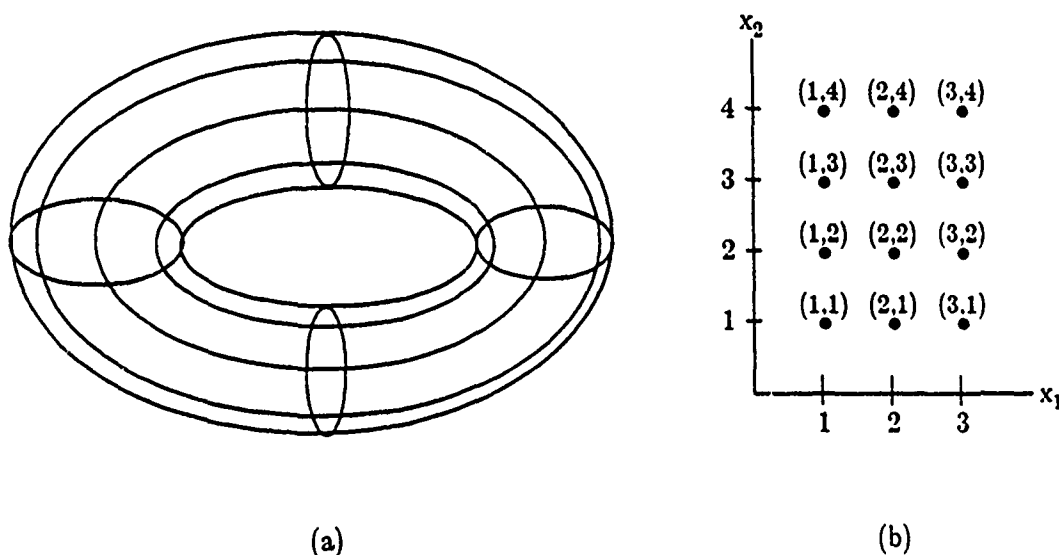


Figure 2. Two Coordinate Sets.

(a) Toroidal Lattice $X \subset \mathbb{R}^3$; (b) A Finite Rectangular Array in \mathbb{R}^2 .

2. A *value set* F is a semi-group. Some value sets we are interested in are the real numbers, the rational numbers, integers, positive reals, positive rationals, and positive integers. These are denoted by \mathbb{R} , \mathbb{Q} , \mathbb{Z} , \mathbb{R}^+ , \mathbb{Q}^+ , and \mathbb{Z}^+ , respectively. We will also be strongly interested in some of the extended number systems. If $F \in \{\mathbb{R}, \mathbb{Q}, \mathbb{Z}, \mathbb{R}^+, \mathbb{Q}^+\}$, then $F_{-\infty}$ denotes $F \cup \{-\infty\}$, $F_{+\infty}$ denotes $F \cup \{+\infty\}$, and $F_{\pm\infty}$ denotes $F \cup \{-\infty, +\infty\}$. We denote an arbitrary value set by F .
3. An F *valued image* \mathbf{a} on a coordinate set X is an element of F^X . Thus, an image $\mathbf{a} \in F^X$ is of form

$$\mathbf{a} = \{(x, a(x)) : x \in X, a(x) \in F\}.$$

4. Let \mathbf{X} and \mathbf{Y} be coordinate sets. An \mathbf{F} -valued template t from \mathbf{Y} to \mathbf{X} is an element of $(\mathbf{F}^{\mathbf{X}})^{\mathbf{Y}}$. For each $y \in \mathbf{Y}$, $t(y)$ is an image on \mathbf{X} . Denoting $t(y)$ by t_y , we have

$$t_y = \{ (x, t_y(x)) : x \in \mathbf{X}, t_y(x) \in \mathbf{F} \} \quad \text{for all } y \in \mathbf{Y}.$$

We give a pictorial representation of a generalized template $t \in (\mathbf{F}^{\mathbf{X}})^{\mathbf{Y}}$ in Figure 3.

They are discussed in detailed in the section below on generalized templates.

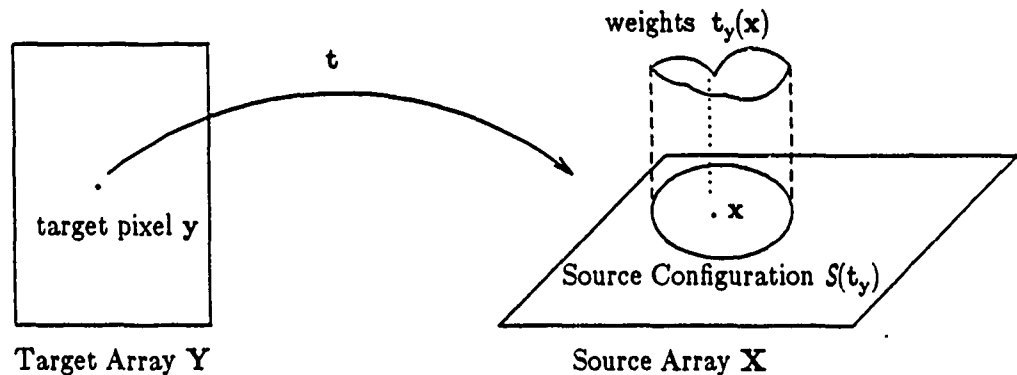


Figure 3. A Pictorial Representation of a Generalized Template.

The set \mathbf{X} is called the set of *image coordinates* of $\mathbf{a} \in \mathbf{F}^{\mathbf{X}}$, and the range of the function \mathbf{a} is called the *image values* of \mathbf{a} . Thus, the image values are a subset of \mathbf{F} . The pair $(x, \mathbf{a}(x))$ is called a *picture element*, or *pixel*, and x is the *pixel location* of the *pixel value* or *gray value* $\mathbf{a}(x)$. We shall use bold lower case letters, \mathbf{x} , to represent a vector in \mathbf{R}^n , and lower case letters (not bold) for the components of the vector. Thus $\mathbf{x} = (x_1, \dots, x_n) \in \mathbf{R}^n$, where $x_i \in \mathbf{R}$ for all i . The set of all \mathbf{F} valued images on \mathbf{X} is denoted by $\mathbf{F}^{\mathbf{X}}$, and the set of all \mathbf{F} valued templates from \mathbf{Y} to \mathbf{X} is denoted by $(\mathbf{F}^{\mathbf{X}})^{\mathbf{Y}}$.

As we will not be using any of the operations concerning coordinate sets or value sets, we refer the reader to other publications discussing this topic [32].

1.1.2. Operations on Images

The basic operations on and between F valued images are the ones induced by the algebraic structure of the value set F . The remaining operations can be defined in terms of these basic ones. In particular, if $F = \mathbb{R}$, then the basic operations for $a, b \in \mathbb{R}^X$ are

$$a + b \equiv \{ (x, c(x)) : c(x) = a(x) + b(x), x \in X \}$$

$$a * b \equiv \{ (x, c(x)) : c(x) = a(x) \cdot b(x), x \in X \}$$

$$a \vee b \equiv \{ (x, c(x)) : c(x) = a(x) \vee b(x), x \in X \}.$$

If X is finite, then we define the *dot product* of two images $a, b \in \mathbb{R}^X$ by

$$a \bullet b = \sum_{x \in X} a(x) \cdot b(x).$$

We say an image $a \in F^X$ is a *constant* image if its gray value at every pixel location is the same. Thus, a constant image $a \in F^X$ has form

$$a(x) = k \in F, \text{ for all } x \in X.$$

In this case we write k for the image a . There are two constant images of importance in the image algebra. One is the *zero* image, defined by $0 \equiv \{ (x, 0) : x \in X \}$, and the *unit* image, defined by $1 \equiv \{ (x, 1) : x \in X \}$. These images have the following properties.

$$a + 0 = 0 + a = a$$

$$a * 1 = 1 * a = a$$

Suppose $f: F \rightarrow F$ is given. Then f induces a function from F^X to F^X , also called f , where

$$f(a) = \{ (x, b(x)) : b(x) = f(a(x)) \}.$$

For example, the function $f: \mathbb{R} \setminus \{0\} \rightarrow \mathbb{R} \setminus \{0\}$ where $f(r) = r^{-1}$ induces a function $f: F^X \rightarrow F^X$, where $f(a) = b$, and $b(x) = 1/a(x)$, if $a(x) \neq 0$, otherwise $b(x) = 0$. The image b so described is denoted by a^{-1} . It is obvious that $a * a^{-1} \neq 1$ for every a . But it is true that

$a * a^{-1} * a = a$. For this reason a^{-1} is called the *pseudo inverse* of a .

If the value set $F = \mathbb{R}_{\pm\infty}$, then the *additive dual* of $\mathbf{a} \in \mathbb{R}_{\pm\infty}^X$ is denoted by \mathbf{a}^* and defined by

$$\mathbf{a}^*(\mathbf{x}) = \begin{cases} -\mathbf{a}(\mathbf{x}) & \text{if } \mathbf{a}(\mathbf{x}) \in \mathbb{R} \\ -\infty & \text{if } \mathbf{a}(\mathbf{x}) = +\infty \\ +\infty & \text{if } \mathbf{a}(\mathbf{x}) = -\infty \end{cases}.$$

Thus we have $(\mathbf{a}^*)^* = \mathbf{a}$.

If $F = \mathbb{R}_{\pm\infty}^+$, then the *multiplicative dual* of $\mathbf{a} \in (\mathbb{R}_{\pm\infty}^+)^X$ is denoted by $\bar{\mathbf{a}}$ and defined by

$$\bar{\mathbf{a}}(\mathbf{x}) = \begin{cases} 1/\mathbf{a}(\mathbf{x}) & \text{if } \mathbf{a}(\mathbf{x}) \in \mathbb{R}^+ \\ -\infty & \text{if } \mathbf{a}(\mathbf{x}) = +\infty \\ +\infty & \text{if } \mathbf{a}(\mathbf{x}) = -\infty \end{cases}.$$

It follows that $\overline{(\bar{\mathbf{a}})} = \mathbf{a}$.

Another useful induced function is the *characteristic function*. Let χ_T denote the usual characteristic function with respect to an arbitrary set T . Here

$$\chi_T(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in T \\ 0 & \text{if } \mathbf{x} \notin T \end{cases}.$$

We now define the *generalized characteristic function* of an image $\mathbf{a} \in \mathbb{R}^X$. Let $\mathbf{a} \in \mathbb{R}^X$ and $S \in (2^F)^X$. Then the *generalized characteristic function* of an image \mathbf{a} is defined as

$$\chi_S(\mathbf{a}) = \mathbf{c} \in \mathbb{R}^X$$

where

$$\mathbf{c} = \{ (\mathbf{x}, \mathbf{c}(\mathbf{x})) : \mathbf{c}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{a}(\mathbf{x}) \in S(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases} \}.$$

Note that the usual characteristic function above is a special case of the generalized characteristic function, where $T \subset F$ and $S(\mathbf{x}) = T$ for all $\mathbf{x} \in X$. The typical thresholding function applied to an image is a simple example of the generalized characteristic function. Fix $\mathbf{b} \in \mathbb{R}^X$. Then $S_{\leq \mathbf{b}} \in (2^{\mathbb{R}})^X$ is defined by

$$S_{\leq b}(x) \equiv \{r \in \mathbf{R} : r \leq b(x)\}.$$

To simplify notation, we define $\chi_{\leq b} \equiv \chi_{S_{\leq b}}$. Thus, for $b, a \in \mathbf{F}^X$, we have

$$\chi_{\leq b}(a) = \{(x, c(x)) : c(x) = \begin{cases} 1 & \text{if } a(x) \leq b(x) \\ 0 & \text{otherwise} \end{cases}\}.$$

If we now consider the characteristic function on $\mathbf{R}_{\pm\infty}^X$, we find that we would like our binary output image to have the values $-\infty$ and 0 instead of 0's and 1's, respectively. We define the *extended* characteristic function as the function induced by

$$\chi_S^\infty(x) = \begin{cases} 0 & \text{if } x \in S \\ -\infty & \text{otherwise} \end{cases}.$$

Thus, $\chi_{\leq b}^\infty(a)$ is defined as

$$\chi_{\leq b}^\infty(a) = \begin{cases} 0 & \text{if } a(x) \leq b(x) \\ -\infty & \text{otherwise} \end{cases}.$$

One unary operation on images is the sum operation, which we will use in Chapter 7. Let X be a finite coordinate set. Then the *sum* of $a \in \mathbf{R}^X$ is defined to be

$$\sum a \equiv a \bullet 1 = \sum_{x \in X} a(x).$$

In context of the lattice structures of $\mathbf{R}_{\pm\infty}$ and $\mathbf{R}_{\pm\infty}^+$, we make the following definition.

Let $a \in \mathbf{R}_{\pm\infty}^X$. The *maximum* of a is the scalar determined by

$$\vee a = \bigvee_{x \in X} a(x).$$

1.1.3. Generalized Templates

For a generalized template $t \in (\mathbf{F}^X)^Y$, the coordinate set Y is called the *target domain* or the domain of t , and X is called the *range space* of t . The pixel location $y \in Y$ at which a template t_y is evaluated is called a *target point* of the template t , and the values $t_y(x)$ are called the *weights* of the template t at y .

If $F \in \{\mathbb{R}, \mathbb{R}_{\pm\infty}, \mathbb{C}\}$, then for $t \in (F^X)^Y$, the set

$$S(t_y) = \{x \in X : t_y(x) \neq 0\}$$

is called the *support of t_y* . If $F \in \{\mathbb{R}_{\pm\infty}, \mathbb{R}_{\pm\infty}^+\}$, then for $t \in (F^X)^Y$ we define

$$S_{-\infty}(t_y) = \{x \in X : t_y(x) \neq -\infty\}$$

and

$$S_{+\infty}(t_y) = \{x \in X : t_y(x) \neq +\infty\}$$

to be the (negative) and positive infinite support, respectively.

If $t \in (F^X)^X$ and for all triples $x, y, z \in X$ with $y + z$ and $x + z \in X$, we have $t_y(x) = t_{y+z}(x+z)$, then t is called *translation invariant*. A template which is not translation invariant is called *translation variant*, or simply *variant*. Translation invariant templates have the nice property that they may be represented pictorially in a concise manner.

The following translation invariant template is presented pictorially in Figure 4. Let $X = Y = Z^2$, and let $y = (i, j) \in Z^2$. Let $x_1 = (i, j)$, $x_2 = (i+1, j)$, $x_3 = (i, j-1)$, and $x_4 =$

$(i+1, j-1)$. Define the weights by $t_y(x) = \begin{cases} i & \text{if } x = x_i, i = 1, \dots, 4 \\ 0 & \text{otherwise} \end{cases}$. Then

$$S(t_y) = \{x_1, \dots, x_4\}.$$

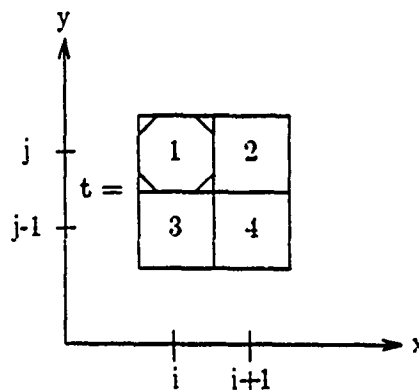


Figure 4. Example of a Translation Invariant Template.

The cell with the hash marks in the pictorial representation of t indicates the location of the target point y .

There are several representations of a template that we will be concerned with. One is the transpose of a template. Let $t \in (F^X)^Y$. Then the *transpose* of t is a template $t' \in (F^Y)^X$ defined by $t'_x(y) \equiv t_y(x)$. If $F \in \{R_{\pm\infty}, R_{\pm\infty}^+\}$, then we can introduce a dual template. For $t \in (R_{\pm\infty}^X)^Y$, the *additive dual* of t is the template $t^* \in (R_{\pm\infty}^Y)^X$ defined by

$$t_x^*(y) = \begin{cases} -t_y(x) & \text{if } t_y(x) \in R \\ -\infty & \text{if } t_y(x) = +\infty \\ +\infty & \text{if } t_y(x) = -\infty \end{cases}$$

Similarly, if $t \in ((R_{\pm\infty}^+)^X)^Y$, the *multiplicative dual* of t is the template $\bar{t} \in ((R_{\pm\infty}^+)^Y)^X$ defined by

$$\bar{t}_x(y) = \begin{cases} 1/t_y(x) & \text{if } t_y(x) \in R^+ \\ -\infty & \text{if } t_y(x) = +\infty \\ +\infty & \text{if } t_y(x) = -\infty \end{cases}$$

1.1.4. Operations Between Images and Templates

One common use of templates is to describe some transformation of an input image based on its image values within a subset of the coordinate set X . We first introduce the *generalized product* between an image and a template. Let $X \subset R^n$ be finite, $X = \{x_1, \dots, x_m\}$. Let γ be an associative binary operation on the value set F . Then the *global reduce operation* Γ on F^X induced by γ is defined by

$$\Gamma(a) = \bigwedge_{x \in X} a(x) = a(x_1) \gamma a(x_2) \gamma \dots \gamma a(x_m).$$

where $a \in F^X$. Thus, $\Gamma: F^X \rightarrow F$.

Images and templates are combined by combining appropriate binary operations. Let F_1 , F_2 , and F be three value sets, and suppose $o: F_1 \times F_2 \rightarrow F$ and $\delta: F_2 \times F_1 \rightarrow F$ are binary operations. If γ is an associative binary operation on F , $a \in F_1^X$, and $t \in (F_2^X)^Y$, then the *generalized backward template operation* of a with t (induced by γ and o) is the binary operation $\odot: F_1^X \times (F_2^X)^Y \rightarrow F^Y$ defined by

$$a \odot t \equiv \{(y, b(y)): b(y) = \bigwedge_{x \in X} a(x) o t_y(x), y \in Y\}.$$

If $t \in (F_2^Y)^X$, then the *generalized forward template operation* of a with t is defined as

$$t \odot a \equiv \{(y, b(y)): b(y) = \bigwedge_{x \in X} t_x(y) \delta a(x), y \in Y\}.$$

Note that the input image a is an F_1 valued image on the coordinate set X , and the output image b is an F valued image on the coordinate set Y , regardless of which template operation, forward or backward, is used. Templates can therefore be used to transform an image on one coordinate set and with values in one set to an image on a completely different coordinate set whose values may be entirely different from the original image's.

Only three special cases of the above generalized operation have been investigated in detail, one by Gader [32] and the other two in this dissertation. Future research will certainly discover other useful combinations. These three operations are denoted by \oplus , \boxtimes , and \odot . The operation \oplus is a linear one, and we refer the interested reader to other references for recent research in this area [32,43,44]. The other two operations, \boxtimes and \odot , are non-linear, and investigation of the structure they induce on images and templates is the focus of this dissertation.

Since our main interest concerns the operations \boxtimes and \odot , we will omit the definition \oplus and refer the interested reader to another reference [31]. Let $X \subset \mathbb{R}^n$ be finite and $Y \subset \mathbb{R}^m$. Let $a \in \mathbb{R}_{-\infty}^X$ and $t \in (\mathbb{R}_{-\infty}^X)^Y$. Then the *backward additive max* is defined as

$$a \boxplus t \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in X} a(x) + t_y(x), y \in Y\},$$

where $\bigvee_{x \in X} a(x) + t_y(x) = \max\{a(x) + t_y(x) : x \in X\}$.

For $t \in (\mathbf{R}_{-\infty}^Y)^X$ we define the *forward additive max* transform by

$$t \boxplus a \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in X} a(x) + t_x(y), y \in Y\}.$$

We use the usual extended arithmetic addition $r + -\infty = -\infty + r = -\infty \forall r \in \mathbf{R}_{-\infty}$ to define $a(x) + t_y(x)$ everywhere.

For $a \in \mathbf{R}_{-\infty}^X$ and $t \in ((\mathbf{R}_{-\infty}^+)^X)^Y$ we define the *backward multiplicative max* transform

$$a \boxtimes t \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in X} a(x) \cdot t_y(x), y \in Y\}.$$

The *forward multiplicative max* transform is given by

$$t \boxtimes a \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in X} a(x) \cdot t_x(y), y \in Y\},$$

where $t \in ((\mathbf{R}_{-\infty}^+)^Y)^X$.

Recall that a *lattice-ordered group*, or *l-group*, is a group which is also a lattice. The operation addition (multiplication) on the l-group \mathbf{R} (\mathbf{R}^+) can be extended in a well-defined manner to addition (multiplication) on $\mathbf{R}_{\pm\infty}$ ($\mathbf{R}_{\pm\infty}^+$) by defining

$$x \times -\infty = -\infty \times x = -\infty, x \in G \cup \{-\infty\}$$

$$x \times +\infty = +\infty \times x = +\infty, x \in G \cup \{+\infty\}$$

$$-\infty \times +\infty = +\infty \times -\infty = -\infty$$

where $\times \in \{+, \cdot\}$, depending on whether $G = \mathbf{R}$ or \mathbf{R}^+ , respectively. Of course, the elements $+\infty, -\infty$ have no additive inverse under the operation $+$ or \cdot , and hence $\mathbf{R}_{\pm\infty}$ (or $\mathbf{R}_{\pm\infty}^+$) is no longer a group. This is discussed in detail in section 1.2, where the notion of a *bounded lattice ordered group*, an extension of a lattice-ordered group with extended

arithmetic, is introduced. This provides for the value set $\mathbf{R}_{\pm\infty}$ to be used in the definition of the image-template operation \boxtimes , for example, and the value set $\mathbf{R}_{\pm\infty}^+$ to be used in the definition of the image-template operation \odot .

We remark that for computational as well as theoretical purposes, we can restate the above two convolutions with the new pixel value calculated only over the support of the template t . If $S_{-\infty}(t_y) \neq \emptyset$, then $\bigvee_{x \in X} a(x) + t_y(x) = \bigvee_{x \in S_{-\infty}(t_y)} a(x) + t_y(x)$, and we have

$$a \boxtimes t \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in S_{-\infty}(t_y)} a(x) + t_y(x), y \in Y\}.$$

Similarly, if $S_{-\infty}(t_y) \neq \emptyset$, then $\bigvee_{x \in X} a(x) * t_y(x) = \bigvee_{x \in S_{-\infty}(t_y)} a(x) * t_y(x)$, and

$$a \odot t \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in S_{-\infty}(t_y)} a(x) * t_y(x), y \in Y\}.$$

If in either case $S_{-\infty}(t_y) = \emptyset$, then we define

$$\bigvee_{x \in S_{-\infty}(t_y)} a(x) + t_y(x) \quad \text{or} \quad \bigvee_{x \in S_{-\infty}(t_y)} a(x) * t_y(x) = -\infty.$$

We may therefore restrict our computation of the new pixel value to the infinite support of t_y . This becomes particularly important when considering mapping of transforms to certain types of parallel architectures, as will be discussed in the introductory remarks to Part II, and Chapter 5.

Because of the duality inherent in the two structures $\mathbf{R}_{\pm\infty}$ and $\mathbf{R}_{\pm\infty}^+$ the operations \boxtimes and \odot induce dual image-template operations, called *additive minimum* and *multiplicative minimum*, respectively. They are defined by

$$a \boxtimes t \equiv (t^* \boxtimes a^*)^*$$

and

$$a \odot t \equiv \overline{(t \odot a)}.$$

Equivalently, we have

$$\begin{aligned} \mathbf{a} \boxtimes \mathbf{t} &\equiv \{(y, b(y)) : b(y) = \bigwedge_{x \in X} a(x) +^l t_y(x), y \in Y\} \\ &= \{(y, b(y)) : b(y) = \bigwedge_{x \in S_{+\infty}(t_y)} a(x) +^l t_y(x), y \in Y\} \end{aligned}$$

and

$$\begin{aligned} \mathbf{t} \oslash \mathbf{a} &\equiv \{(y, b(y)) : b(y) = \bigwedge_{x \in X} a(x) *^l t_x(y), y \in Y\} \\ &= \bigwedge_{x \in S_{+\infty}(t_x)} a(x) *^l t_x(y), y \in Y. \end{aligned}$$

where the dual operations $+^l$ and $*^l$ are presented in section 1.2. As before, if $S_{+\infty}(t_y) = \emptyset$, we define

$$\bigwedge_{x \in S_{+\infty}(t_x)} a(x) +^l t_x(y) \quad \text{or} \quad \bigwedge_{x \in S_{+\infty}(t_x)} a(x) *^l t_x(y) = +\infty.$$

The above definitions assume that the support $S_{-\infty}(t_y)$ is finite for each $y \in Y$. We may extend the above definitions to continuous functions \mathbf{a} and \mathbf{t}_y on a compact set $S_{-\infty}(t_y)$. This is well-defined as the sum or product of two continuous functions on a compact subset of \mathbb{R}^n , which is continuous, always contains a maximum. Extending the basic properties of the image algebra operations involving \boxtimes and \oslash from the discrete case to the continuous case should present little difficulty, and remains an open problem at this time.

1.1.5. Operations Between Generalized Templates

The pointwise operations of the value set F can also be extended to to operations between templates. For example, if $F = \mathbb{R}$, then we have

$$\mathbf{s} + \mathbf{t} \equiv \mathbf{r}, \text{ where } r_y = s_y + t_y$$

$$\mathbf{s} * \mathbf{t} \equiv \mathbf{r}, \text{ where } r_y = s_y * t_y$$

$$\mathbf{s} \vee \mathbf{t} \equiv \mathbf{r}, \text{ where } r_y = s_y \vee t_y.$$

If $F = R_{\pm\infty}$ then we define

$$s + t \equiv r, \text{ where } r_y(x) = \begin{cases} s_y(x) + t_y(x) & \text{if } x \in S_{-\infty}(t_y) \cap S_{-\infty}(s_y) \\ s_y(x) & \text{if } x \in S_{-\infty}(s_y) \setminus S_{-\infty}(t_y) \\ t_y(x) & \text{if } x \in S_{-\infty}(t_y) \setminus S_{-\infty}(s_y) \\ -\infty & \text{otherwise} \end{cases}$$

Note that in the case where s and t have no values of $-\infty$ or $+\infty$ anywhere, then the definition of $s + t$ on the value set $R_{\pm\infty}$ degenerates to the definition of $s + t$ on the value set R .

The generalized image-template operation \oplus generalizes to a generalized template-template product. Let $X \subset R^n$ be finite, $X = \{x_1, \dots, x_m\}$, and let γ be an associative binary operation on the value set F with global reduce operation Γ on F^X . Let F_1, F_2 , and F be three value sets, and suppose $\circ : F_1 \times F_2 \rightarrow F$ is a commutative binary operation. If γ is an associative binary operation on F , $t \in (F_1^X)^W$, and $s \in (F_2^W)^Y$, then the *generalized template operation of t with s (induced by γ and \circ)* is the binary operation $\oplus : (F_1^X)^W \times (F_2^W)^Y$ defined by

$$t \oplus s \equiv r \in (F^X)^Y, \text{ where}$$

$$r_y(x) = \Gamma_{w \in W} t_w(x) \circ s_y(w), y \in Y, x \in X.$$

Note that if $|X| = 1$, then the definition of the generalized template operation of t and s degenerates to the definition of the generalized backward template operation of the image $t \in F_1^W$ with the template $s \in (F_2^W)^Y$, and $r \in F^Y$. If $|Y| = 1$, then the definition of the generalized template operation of t and s degenerates to the definition of the forward template operation of the image $s \in F_2^W$ with the template $t \in (F_1^X)^W$, where $r \in F^X$.

The specific cases for $\oplus = \oplus$, \boxtimes , or \odot thus generalize to operations between templates. We give the definitions for \boxtimes and \odot , and refer the reader to another reference for

the definition of \oplus [32]. Let $t \in (R_{\pm\infty}^X)^Y$ and $s \in (R_{\pm\infty}^W)^X$. Then $s \boxtimes t = r \in (R_{\pm\infty}^W)^Y$ is defined by

$$r_y(w) = \bigvee_{x \in X} t_y(x) + s_x(w), \text{ where } w \in W.$$

Again, as in the image-template operations, we may restrict our computation to a subset of X . In particular, for $y \in Y$, we define the set

$$S_{-\infty}(w) = \{x \in X : x \in S_{-\infty}(t_y) \text{ and } w \in S_{-\infty}(s_x)\}.$$

Then $r = s \boxtimes t \in (R_{\pm\infty}^W)^Y$ is defined by

$$r_y(w) = \bigvee_{x \in S_{-\infty}(w)} t_y(x) + s_x(w),$$

where we define $\bigvee_{x \in S_{-\infty}(w)} t_y(x) + s_x(w) = -\infty$ whenever $S_{-\infty}(w) = \emptyset$.

The operation \odot has a similar situation. We have $r = s \odot t \in ((R_{\pm\infty}^+)^W)^Y$ which is defined by

$$r_y(w) = \bigvee_{x \in S(w)} t_y(x) \cdot s_x(w),$$

where we define $\bigvee_{x \in S(w)} t_y(x) \cdot s_x(w) = -\infty$ whenever $S(w) = \emptyset$.

It follows from these definitions that the infinite support of the template r is

$$S_{-\infty}(r_y) = \{w \in W : S_{-\infty}(w) \neq \emptyset\}.$$

The definitions given in this section are the elemental ones. Further definitions that play important parts in the theoretical development of the lattice structure of the image algebra will be presented as needed.

We define the complementary operations of \boxtimes and \odot for templates in the natural way. Let $t \in (R_{\pm\infty}^X)^Y$ and $s \in (R_{\pm\infty}^W)^X$. Then $s \boxtimes t \in (R_{\pm\infty}^W)^Y$ is defined by

$$s \boxtimes t \equiv (t^* \boxtimes s^*)^*$$

Similarly, for $t \in ((R_{\pm\infty}^+)^X)^Y$ and $s \in ((R_{\pm\infty}^+)^W)^X$, $s \otimes t \in ((R_{\pm\infty}^+)^W)^Y$ is defined by

$$s \otimes t \equiv \overline{(t \otimes s)}.$$

We would like to remark upon one notational deviation between the *Overview's* [32] definition for the \otimes operations and the one presented here. Let $R_{+\infty}^{\geq 0} = \{r \in \mathbf{R} : r \geq 0\} \cup \{+\infty\}$. In the *Overview*, for $a \in (R_{+\infty}^{\geq 0})^X$ and $t \in ((R_{+\infty}^{\geq 0})^X)^Y$, the *backward multiplicative max transform* is defined as

$$a \otimes t \equiv \{(y, b(y)) : b(y) = \bigvee_{x \in X} a(x) \cdot t_y(x), y \in Y\}$$

which is equivalent to

$$a \otimes t = \{(y, b(y)) : b(y) = \bigvee_{x \in S(t_y)} a(x) \cdot t_y(x), y \in Y\}$$

with $b(y) = 0$ if $S(t_y) = \emptyset$. The difference between the definition given earlier and this one is the *value set*, namely $R_{\pm\infty}^+$ in this document and $R_{+\infty}^{\geq 0}$ in the *Overview*. The number 0 acts as a lower bound in $R_{+\infty}^{\geq 0}$ exactly as $-\infty$ acts as a lower bound in $R_{\pm\infty}^+$. Multiplication of the element 0 with the element ∞ follows the same rules as multiplication of the element $-\infty$ with the element ∞ as given on page 18. In other words, the element 0 can replace *symbolically* the element $-\infty$. The main advantage of using the number 0 instead of the symbol $-\infty$ is for ease of machine and software implementation. Most real image processing data will have no values corresponding to the symbol $+\infty$, and quite often have non-negative values, including 0's. Using 0 as the bottom element enables that value to be represented easily in the computer, while special programming methods would have to be considered to represent the symbol $-\infty$. For purposes which will become clear in the course of this document, we have remained with the notation $R_{\pm\infty}^+$. In implementing any of the ideas in this dissertation, if the value set at hand is $R_{\pm\infty}^+$ it should be clear that the symbol $-\infty$ can be replaced with a 0 and $S_{-\infty}(t(y))$ replaced by $S(t(y))$, so that representation in computers may be more easily accomplished.

1.2 Minimax Algebra

The last 40 years have seen a number of different authors discover, apparently independently, a non-linear algebraic structure, which each has used to solve a different type of problem. The operands of this algebra are the real numbers, with $-\infty$ (or $+\infty$ adjoined), with the two binary operations of addition and maximum (or minimum). The extension of this structure to matrices was formalized mathematically, in the environment in which the above problems were posed, by Cuninghame-Green in his book *Minimax Algebra* [39]. It is well known that the structure of \mathbf{R} with the operations of $+$ and \vee is a semi-lattice ordered group, and that $(\mathbf{R}, \vee, \wedge, +)$ is a lattice-ordered group, or an l-group [36]. Viewing $\mathbf{R}_{-\infty}$ as a set with the two binary operations of $+$ and \vee , and then investigating the structure of the set of all $n \times n$ matrices with values in $\mathbf{R}_{-\infty}$, leads to an entirely different perspective of a class of non-linear operators. These ideas were applied by Shimbel [46] to communications networks. Two authors, Cuninghame-Green [37,38] and Giffler [47] applied them to the problem of machine-scheduling. Others [48,49,50,51] have discussed their usefulness in applications to shortest path problems in graphs. Cuninghame-Green gives several examples throughout his book [39], primarily in the field of operations research. Another useful application, to image algebra, was again independently developed by G.X Ritter et al. [52].

In fact, the notion of a matrix product can be generalized to what is called the *generalized matrix product* [40], whose definition is given below.

Let \mathbf{F} denote a set of numbers. Let f and g be functions from $\mathbf{F} \times \mathbf{F}$ into \mathbf{F} . For simplicity, assume the binary operation f to be associative. Let \mathbf{F}^{mp} denote the set of all $m \times p$ matrices with values in \mathbf{F} , and let $(a_{ij}) = A \in \mathbf{F}^{mp}$ and $(b_{jk}) = B \in \mathbf{F}^{pn}$.

Define $f \cdot g$ to be the function from $\mathbf{F}^{mp} \times \mathbf{F}^{pn}$ into \mathbf{F}^{mn} given by

$$(f \cdot g)(A, B) = C,$$

where $c_{ik} = (a_{i1} g b_{1k}) f (a_{i2} g b_{2k}) f \cdots f (a_{ip} g b_{pk})$, for $i = 1, \dots, m$, $k = 1, \dots, n$, and f and g are viewed as binary operations.

Thus, if f denotes addition and g multiplication, then $(f \cdot g)(A, B)$ is the ordinary matrix product of matrices A and B . Cuninghame-Green develops the setting for a formal matrix calculus based on the two binary operations $+$ and \vee of the extended real numbers, analogous to linear algebra which uses the two operations of multiplication and arithmetic sum. He terms this matrix theory *minimax matrix theory*. The development of the theory is performed in the abstract, with an eye towards applications for matrices with values in the set $\mathbf{R}_{\pm\infty}$. The importance of Cuninghame-Green's work to the image algebra is that not only is the minimax matrix algebra embedded in the image algebra for the set $\mathbf{R}_{\pm\infty}$ but also for the set $\mathbf{R}_{\pm\infty}^+$. The set $(\mathbf{R}^+, \vee, \wedge, *)$ is an l-group also. An image algebra transform using either \boxtimes or \odot can thus be viewed as a matrix transform in the minimax algebra for the respective case of $\mathbf{R}_{\pm\infty}$ or $\mathbf{R}_{\pm\infty}^+$. This completes the mathematical identification of the three main subalgebras in the image algebra. The linear transforms were classified by Gader [33] who showed that linear algebra is embedded into image algebra. As a result of each embedding above, the full power of the respective mathematical theory can be applied to solving problems in image processing, as long as the image processing problem can be formulated using image algebra operations of \oplus , \boxtimes , or \odot . Since it has been formally proven that the image algebra can represent all image-to-image transforms (see section 1.1), the embeddings are very useful to have.

The rest of this section is devoted to introducing the basic notions of the minimax algebra structure and properties.

1.2.1. Basic Definitions and Notation

Let F be a semi-lattice ordered semi-group with semi-lattice operation \vee and semi-group operation \times . Thus, F satisfies

$$x \vee (y \vee z) = (x \vee y) \vee z \quad A_1$$

$$x \vee y = y \vee x \quad A_2$$

$$x \vee x = x \quad A_3$$

as it is a semi-lattice, as well as

$$x \times (y \times z) = (x \times y) \times z \quad A_4$$

as it has an associative group operation \times , and

$$x \times (y \vee z) = (x \times y) \vee (x \times z) \quad A_5$$

$$(y \vee z) \times x = (y \times x) \vee (z \times x) \quad A_6$$

as it is an ordered semi-group. We call this structure a *belt*, in the vein of rings. The operation \vee is called an *addition*, and the operation \times a *multiplication*. We shall also call a semi-lattice an *s-lattice*.

Suppose the belt F also satisfies the *dual* to axioms A_1 through A_6 , where \times' is another binary group multiplication:

$$x \wedge (y \wedge z) = (x \wedge y) \wedge z \quad A'_1$$

$$x \wedge y = y \wedge x \quad A'_2$$

$$x \wedge x = x \quad A'_3$$

$$x \times' (y \times' z) = (x \times' y) \times' z \quad A'_4$$

$$x \times' (y \wedge z) = (x \times' y) \wedge (x \times' z) \quad A'_5$$

$$(y \wedge z) \times' x = (y \times' x) \wedge (z \times' x). \quad A'_6$$

Here, \times' is called a *dual multiplication*, and \wedge is called a *dual addition*. The (group) multiplication or dual multiplication is not assumed to be commutative.

If in addition to the above 12 axioms, F satisfies the following axiom,

$$x \vee (y \wedge x) = x \wedge (y \vee x) = x,$$

then F is a *belt with duality*. If the multiplication \times and dual multiplication \times' coincide, then we call the multiplication *self-dual*. A belt with duality and self-dual multiplication corresponds to a *lattice-ordered semi-group*, or *l-semi-group*, in lattice theory.

Let (F_1, \vee) and (F_2, \vee) be two *s-lattices*. A function $f: F_1 \rightarrow F_2$ is an *s-lattice homomorphism* if

$$f(x \vee y) = f(x) \vee f(y),$$

for all $x, y \in F$. If F_1 and F_2 are belts and $f: F_1 \rightarrow F_2$ is an *s-lattice homomorphism*, then if f also satisfies

$$f(x \times y) = f(x) \times f(y)$$

for all $x, y \in F$, then we say that f is a *belt homomorphism*. The following is an example of a belt *isomorphism*. Define $f: \mathbf{R} \rightarrow \mathbf{R}^+$ by

$$f(x) = e^x.$$

Then $f(x \vee y) = f(x) \vee f(y)$, and $f(x + y) = f(x) * f(y)$. It is trivial to show that f is a belt isomorphism.

The belts \mathbf{R} and \mathbf{R}^+ are *commutative belts*, that is, the multiplication \times commutes. Each also has an *identity element* under the multiplication, namely 0 for \mathbf{R} and 1 for \mathbf{R}^+ . Because they are groups, each element $r \in F$ has a unique multiplicative inverse; we call such a belt a *division belt*, by analogy with division rings. A belt has a *null element* if there exists an element $\theta \in F$ such that

$$\forall x \in F, x \vee \theta = x \text{ and } x \times \theta = \theta \times x = \theta.$$

The belts $(\mathbf{R}_{-\infty}, \vee, +)$ and $(\mathbf{R}_{-\infty}^+, \vee, +)$ each have the element $-\infty$ as its null element.

A division belt with distinct operations \times and \vee and with duality corresponds to a *lattice-ordered group*, or *l-group*. In fact, if (F, \vee, \times) is a belt with distinct operations \vee and \times , then by defining

$$x \wedge y = (x^{-1} \vee y^{-1})^{-1}, \quad \forall x, y \in F \quad (1-2)$$

we have introduced a second (dual) s-lattice operation \wedge such that (F, \vee, \wedge) becomes a (distributive) lattice [36]. In our terms, the division belt F acquires a duality with a self-dual multiplication. Our main interest will be for the l-groups $(F, \vee, \times, \wedge, \times') = (R, \vee, +, \wedge, +)$ and $(R^+, \vee, *, \wedge, *)$, $*$ representing real multiplication. From the above discussion, it follows that $(R, \vee, +, \wedge, +)$ and $(R^+, \vee, *, \wedge, *)$ are isomorphic as l-groups.

An arbitrary l-group F having two distinct binary operations \vee and \times can be extended in the following way. We adjoin the elements $+\infty$ and $-\infty$ to the set F and denote this new set by $F_{\pm\infty}$, where $-\infty < x < +\infty \quad \forall x \in F$. We define a multiplication and a dual multiplication in $F_{\pm\infty}$ by: if $x, y \in F$, then $x \times y$ is already defined. Otherwise,

$$x \times -\infty = -\infty \times x = -\infty, \quad x \in F \cup \{-\infty\}$$

$$x \times +\infty = +\infty \times x = +\infty, \quad x \in F \cup \{+\infty\}$$

$$x \times' -\infty = -\infty \times' x = -\infty, \quad x \in F \cup \{-\infty\}$$

$$x \times' +\infty = +\infty \times' x = +\infty, \quad x \in F \cup \{+\infty\}.$$

$$-\infty \times +\infty = +\infty \times -\infty = -\infty$$

$$-\infty \times' +\infty = +\infty \times' -\infty = +\infty$$

The element $-\infty$ acts as a null element in the entire system $(F_{\pm\infty}, \vee, \times)$ and the element $+\infty$ acts as a null element in the entire system $(F_{\pm\infty}, \wedge, \times')$. However, the multiplications \times and \times' are asymmetric between the elements $-\infty$ and $+\infty$. The elements in F are called the *finite elements*.

We call such a system $(F_{\pm\infty}, V, \times, \wedge, \times')$ a *bounded l-group*, and F is called the *group* of the bounded l-group $F_{\pm\infty}$.

The two bounded l-groups $(R_{\pm\infty}, V, +, \wedge, +')$ and $(R_{\pm\infty}^+, V, *, \wedge, *')$ will be our main concern. Another bounded l-group of interest is the *3-element* bounded l-group with group ϕ , denoted by F_3 . Note that the boolean algebra $(\{-\infty, \phi\}, V, \wedge)$ is embedded in F_3 , with $OR = V$ (maximum), $AND = \wedge$ (minimum), $FALSE = -\infty$, and $TRUE = \phi$. It is simple to check that the familiar truth tables hold.

Let (F, V, \times) be a belt, and let (T, V) be an s-lattice. Suppose we have a right multiplication of elements of T by elements of F :

$$x \times \lambda \in T \quad \forall \text{ pairs } x, \lambda, x \in T, \lambda \in F.$$

We call (T, V) a *right s-lattice space over (F, V, \times)* , or just say T is a *space over F* if the following axioms are satisfied for all $x, y \in T$ and for all $\lambda, \mu \in F$:

$$(T, V) \text{ is an s-lattice}$$

$$(x \times \lambda) \times \mu = x \times (\lambda \times \mu)$$

$$(x \vee y) \times \lambda = (x \times \lambda) \vee (y \times \lambda)$$

$$x \times (\lambda \vee \mu) = (x \times \lambda) \vee (x \times \mu)$$

and if F has an identity element ϕ ,

$$x \times \phi = x.$$

Such spaces play the role of vector spaces in the minimax theory. If T and F are known, then we shall simply say that T is a *space*.

A *subspace* is a subset of a space which is itself a space over the belt F .

Let $(S, V), (T, V)$ be given spaces over a belt (F, V, \times) . An s-lattice homomorphism

$g: (S, V) \rightarrow (T, V)$ is called *right linear (over F)* if

$$g(x \times \lambda) = g(x) \times \lambda \quad \forall x \in S, \forall \lambda \in F.$$

We denote the set of all right-linear homomorphisms from S to T over F by $\text{Hom}_F(S, T)$. That is,

$$\text{Hom}_F(S, T) = \{g: S \rightarrow T \text{ is a homomorphism and } g(x \times \lambda) = g(x) \times \lambda \quad \forall x \in S, \forall \lambda \in F\}.$$

Let (F, V, \times) be a belt and (T, V) be an s -lattice, and suppose we have defined a left multiplication of elements of T by elements of F :

$$\lambda \times x \in T \quad \forall \text{ pairs } x, \lambda, x \in T, \lambda \in F.$$

The left variants of the above five axioms are easily stated. We define a system satisfying those left axioms a *left space over F*. This allows us to define a two-sided space. A *two-sided space* is a triple (L, T, R) such that

L is a belt and T is a left space over L .

R is a belt and T is a right space over R .

$$\forall \lambda \in L, \forall x \in T \text{ and } \forall \mu \in R: \lambda \times (x \times \mu) = (\lambda \times x) \times \mu.$$

Let (F, V, \times) be a belt. An important class of spaces over F is the class of function spaces. Here, the s -lattice (T, V) is (F^U, V) . Such spaces are naturally two-sided. We shall only be interested in the case where $|U| = n \in \mathbb{Z}^+$. A space (T, V) is of form (F^n, V) , and hence our spaces F^n are spaces of n -tuples.

When discussing conjugacy in linear operator theory, two approaches are commonly used. One defines the conjugate of a given space S as a special set S^* of linear, scalar-valued functions defined on S . The other involves defining an *involution* taking $x \in S$ to $x^* \in S^*$ which satisfy certain axioms. (Recall a function f is an involution if $f^{-1}(f(x)) = x$.) The situation is slightly more complicated in the case of lattice transforms.

Let (S, V, \times) and (T, \wedge, \times') be given belts. We say that (T, \wedge, \times') is *conjugate* to (S, V, \times) if there is a function $g: S \rightarrow T$ such that

$$g \text{ is bijective} \quad C_1$$

$$\forall x, y \in S, g(x \vee y) = g(x) \wedge g(y) \quad C_2$$

$$\forall x, y \in S, g(x \times y) = g(y) \times' g(x). \quad C_3$$

In lattice theory, g is called a *dual isomorphism*. Note that conjugacy is a symmetric relation. If (S, V, \wedge) is an s -lattice with duality satisfying the first two axioms, then we say that S is *self-conjugate*. If $(S, V, \times, \wedge, \times')$ a belt with duality, we say that $(S, V, \times, \wedge, \times')$ is *self-conjugate* if (S, \wedge, \times') is conjugate to (S, V, \times) .

In particular, every division belt is self-conjugate under the bijection $x^* = x^{-1}$, and every bounded l -group is self-conjugate under the bijection $(-\infty)^* = +\infty$, $(+\infty)^* = -\infty$, and $x^* = x^{-1}$ if x is finite.

1.2.2. Matrix Algebra

We now present the extension of the belt operations to matrices. Let (F, V, \times) be a belt. Let M_{mn} be the set of all $m \times n$ matrices with values in the set F , and let $s = (s_{ij}), t = (t_{ij}) \in M_{mn}$. Then we define

$$(s_{ij}) \vee (t_{ij}) \equiv (s_{ij} \vee t_{ij})$$

and for $(s_{ij}) \in M_{mh}, (t_{jk}) \in M_{hn}$, we have

$$(s_{ij}) \times (t_{jk}) \equiv \left(\bigvee_{j=1}^h [s_{ij} \times t_{jk}] \right) \in M_{mn}.$$

Suppose $s \in M_{mn}$ and $t \in M_{hq}$. We say that s and t are *conformable for addition* whenever both $m = h$ and $n = q$, and *conformable for multiplication* whenever $n = h$. For the remainder of this presentation, we use the notation F^n and M_{mn} , as defined above. Also, we call an n -tuple or a matrix *finite* if all its elements are finite, i.e. not equal to either $+\infty$ or $-\infty$.

If $(F, V, \times, \wedge, \times')$ is a belt with duality, then we say that a space (T, V) over F has a *duality* if

a dual addition \wedge is defined where (T, V, \wedge) is an s -lattice with duality;

(T, \wedge) is a space over the belt (F, \wedge, \times') .

We also have a dual matrix addition and dual multiplication defined for matrices over a belt with duality.

$$(s_{ij}) \wedge (t_{ij}) \equiv (s_{ij} \wedge t_{ij})$$

and for $(s_{ij}) \in M_{mh}, (t_{jk}) \in M_{hn}$, we have

$$(s_{ij}) \times' (t_{jk}) \equiv \left(\bigwedge_{j=1}^h [s_{ij} \times' t_{jk}] \right) \in M_{mn}$$

with the expressions *conformable for dual addition* \wedge and *conformable for dual multiplication* \times' used in the obvious way.

Let (F, V, \times) be a belt and let M_{pq} denote the set of $p \times q$ matrices with values in F .

The following are some basic properties that are proven in [39].

- (1) (M_{mn}, V) is an s -lattice and (M_{np}, V) is a function space over (F, V, \times) ;
- (2) (M_{nn}, V, \times) is a belt;
- (3) (M_{np}, V) is a left space over the belt (M_{nn}, V, \times) ;
- (4) M_{np} is a right space over the belt F ;
- (5) Scalar multiplication of a matrix s by an element $\lambda \in F$ is defined by

$$(s_{ij}) \times \lambda \equiv (s_{ij} \times \lambda)$$

$$\lambda \times (s_{ij}) \equiv (\lambda \times s_{ij})$$

for all $(s_{ij}) \in M_{np}, \lambda \in F$;

- (6) For all $s \in M_{mn}, t, u \in M_{np}, \lambda \in F$,

$$\begin{aligned} s \times (t \vee u) &= (s \times t) \vee (s \times u) \\ s \times (t \times \lambda) &= (s \times t) \times \lambda. \end{aligned}$$

Since the s -lattice (M_{nl}, \vee) is isomorphic to the s -lattice F^n , we have F^n is a function space over F as well as a space over M_{nn} . This mimics the classical role of matrices as linear transformations of spaces of n -tuples!

Two important matrices in our present setting are the identity matrix and the null matrix. Suppose the belt F has identity and null elements ϕ and $-\infty$ respectively. We define the *identity matrix* $e \in M_{nn}$ by

$$e = \begin{bmatrix} \phi & . & . & . & . \\ . & \phi & . & -\infty & . \\ . & . & . & . & . \\ . & -\infty & . & . & . \\ . & . & . & . & \phi \end{bmatrix}$$

and the *null matrix* $\Phi \in M_{nn}$ by

$$\Phi = \begin{bmatrix} -\infty & . & . & . & . \\ . & -\infty & . & -\infty & . \\ . & . & . & . & . \\ . & -\infty & . & . & . \\ . & . & . & . & -\infty \end{bmatrix}.$$

Thus we have $\forall s \in M_{nn}$ and for $\Phi \in M_{nn}$,

$$e \times s = s \times e = s$$

$$s \vee \Phi = s$$

$$s \times \Phi = \Phi \times s = \Phi.$$

In the bounded l -group $R_{\pm\infty}$ we have

$$e = \begin{bmatrix} 0 & . & . & . & . \\ . & 0 & . & -\infty & . \\ . & . & . & . & . \\ . & -\infty & . & . & . \\ . & . & . & . & 0 \end{bmatrix}$$

and in $R_{\pm\infty}^+$ we have

$$e = \begin{bmatrix} 1 & . & . & . & . \\ . & 1 & . & -\infty & . \\ . & . & . & . & . \\ . & -\infty & . & . & . \\ . & . & . & . & 1 \end{bmatrix}.$$

Conjugacy extends to matrices if the underlying value set is itself a self-conjugate belt. This is stated in the next proposition.

Proposition 1.1 [39]. *If $(F, \vee, \times, \wedge, \times')$ is a self-conjugate belt, then $(M_{nn}, \vee, \times, \wedge, \times')$ is a self-conjugate belt.*

In linear algebra, we characterize linear transformations of vector spaces entirely in terms of matrices. Are we able to do a similar classification here? The following results give necessary and sufficient conditions for this to be the case.

Theorem 1.2 [39]. *Let F be a belt which has an identity element ϕ with respect to \times and a null element θ with respect to $-\infty$. Then for all integers $m, n \geq 1$, M_{mn} is isomorphic to $\text{Hom}_F(F^n, F^m)$.*

Corollary 1.3 [39]. *Let F be a belt, and let $n > 1$ be a given integer. Then a necessary and sufficient condition that M_{mn} be isomorphic to $\text{Hom}_F(F^n, F^m)$ for all integers $n, m \geq 1$ is that F have an identity element ϕ with respect to \times and a null element θ with respect to \vee .*

We call a matrix $s \in M_{mn}$ a *lattice transform*.

Many of the results that were stated in Cuninghame's book can be viewed in context of a dual lattice-ordered semi-group, which has been extensively researched [36]. However, we wish to study the structure from a different perspective. The extension of the belt operations to matrices allows us to view matrices as operators on spaces of n -tuples, in a way similar to vector-space transformations. These operators are non-linear due to the lattice

structure of the underlying set F . Thus, we may study this particular class of non-linear transforms in a mathematically rigorous setting, and, since an image can be viewed as a vector and a template as a matrix (as will be shown in Chapter 2), apply results from the minimax matrix theory directly to solve image processing problems. For example, decomposition of matrices corresponds to decomposition of templates. This particular application is discussed in Chapter 5.

CHAPTER 2 THE ISOMORPHISM

In his Ph.D. dissertation, P. Gader showed that linear algebra can be embedded into the image algebra [32]. One very powerful implication of this is that all the tools of linear algebra are directly applicable to solving problems in image processing whenever the image algebra operation \oplus is involved. We now show an embedding of the minimax algebra into image algebra for the two cases where the belts are \mathbf{R} and \mathbf{R}^+ . We employ the same functions Ψ and ν as used by Gader in his dissertation.

Let \mathbf{X} and \mathbf{Y} be finite arrays, with $|\mathbf{X}| = m$ and $|\mathbf{Y}| = n$. Assume the points of \mathbf{X} are labelled lexicographically $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$. Assume a similar labelling for \mathbf{Y} : $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$. Let $\mathbf{R}_{\pm\infty}$ have its usual meaning. Let $\mathbf{R}_{\pm\infty}^m = \{(x_1, x_2, \dots, x_m) : x_i \in \mathbf{R}_{\pm\infty}\}$. That is, $\mathbf{R}_{\pm\infty}^m$ is the set of row vectors of m -tuples with values in $\mathbf{R}_{\pm\infty}$. Let $\mathbf{a} \in \mathbf{R}_{\pm\infty}^{\mathbf{X}}$, M_{mn} denote the set of $m \times n$ matrices with values in $\mathbf{R}_{\pm\infty}$ and define $\nu : \mathbf{R}_{\pm\infty}^{\mathbf{X}} \rightarrow \mathbf{R}_{\pm\infty}^m$ by

$$\nu(\mathbf{a}) = (\mathbf{a}(\mathbf{x}_1), \dots, \mathbf{a}(\mathbf{x}_m)).$$

Define $\Psi : (\mathbf{R}_{\pm\infty}^{\mathbf{X}})^{\mathbf{Y}} \rightarrow M_{mn}$ by

$$\Psi(\mathbf{t}) = M_{\mathbf{t}} = (p_{ij}), \text{ where } p_{ij} = t_{y_j}(\mathbf{x}_i).$$

Note that the j -th column of $M_{\mathbf{t}}$ is simply $(\nu(\mathbf{t}_j))'$, the prime denoting transpose.

In the following lemmas, we assume that $|\mathbf{X}| = m$, $|\mathbf{Y}| = n$, and $|\mathbf{W}| = l$. We claim the following:

Lemma 2.1. $\nu(\mathbf{a} \boxtimes \mathbf{t}) = \nu(\mathbf{a}) \times \Psi(\mathbf{t})$, for $\mathbf{t} \in (\mathbf{R}_{\pm\infty}^{\mathbf{X}})^{\mathbf{Y}}$, $\mathbf{a} \in \mathbf{R}_{\pm\infty}^{\mathbf{X}}$.

Lemma 2.2. $\nu(a \vee b) = \nu(a) \vee \nu(b)$, $a \in F_{\pm\infty}^X$, $F \in \{R, R^+\}$.

Lemma 2.3. $\Psi(s \boxtimes t) = \Psi(s) \times \Psi(t)$, for $s \in (R_{\pm\infty}^X)^W$, $t \in (R_{\pm\infty}^W)^Y$.

Lemma 2.4. $\Psi(s \vee t) = \Psi(s) \vee \Psi(t)$, $s, t \in (F_{\pm\infty}^X)^Y$, $F \in \{R, R^+\}$.

The proofs are given below.

Proof to Lemma 2.1.

We must show that

$$(a \boxtimes t)(y_k) = (\nu(a) \times \Psi(t))_k.$$

First note that $\nu(a \boxtimes t)$ is a $1 \times n$ row vector, as is $\nu(a) \times \Psi(t)$. We have

$$(a \boxtimes t)(y_k) = \bigvee_{x \in X} a(x) + t_{y_k}(x) = \bigvee_{i=1}^m a(x_i) + t_{y_k}(x_i).$$

$$\text{Also, } (\nu(a) \times \Psi(t))_k = \bigvee_{j=1}^m (\nu(a))_j + (\Psi(t))_{jk} = \bigvee_{j=1}^m a(x_j) + t_{y_k}(x_j).$$

Q.E.D.

Proof to Lemma 2.2.

At location x_k , the image $a \vee b$ has value $a(x_k) \vee b(x_k)$. At location k , the row vector $\nu(a) \vee \nu(b)$ has value $(\nu(a))_k \vee (\nu(b))_k = a(x_k) \vee b(x_k)$.

Q.E.D.

Proof to Lemma 2.3.

Here, $s \in (R_{\pm\infty}^X)^W$ and $t \in (R_{\pm\infty}^W)^Y$ implies

$$s \boxtimes t = r \in (R_{\pm\infty}^X)^Y,$$

and

$$r_{y_j}(x_i) = \bigvee_{w \in W} t_{y_j}(w) + s_w(x_i) = \bigvee_{k=1}^l t_{y_j}(w_k) + s_{w_k}(x_i).$$

Now, let $\Psi(s) \times \Psi(t) = u \in M_{mn}$. We have

$$u_{ij} = \bigvee_{k=1}^I (\Psi(s))_{ik} + (\Psi(t))_{kj} = \bigvee_{k=1}^I s_{w_k}(x_i) + t_{y_j}(w_k) = \bigvee_{k=1}^I t_{y_j}(w_k) + s_{w_k}(x_i).$$

Q.E.D.

Proof to Lemma 2.4.

Here, $s, t \in (R_{\pm\infty}^X)^Y$. Then

$$(s \vee t)_{y_j}(x_i) = s_{y_j}(x_i) \vee t_{y_j}(x_i),$$

while

$$(\Psi(s) \vee \Psi(t))_{ij} = (\Psi(s))_{ij} \vee (\Psi(t))_{ij} = s_{y_j}(x_i) \vee t_{y_j}(x_i).$$

Q.E.D.

In order to prove the isomorphism theorem, we will use the following lemma.

Lemma 2.5. $\Psi(t^*) = (\Psi(t))^*$, $t \in (F_{\pm\infty}^X)^Y$, where F denotes either R or R^+ . In this particular instance we let t^* denote the conjugate template of $t \in (F_{\pm\infty}^X)^Y$.

Proof: Let $s = t^*$. Then $s \in (F_{\pm\infty}^Y)^X$, and

$$\Psi(t^*) = \Psi(s) = M_s = (p_{ij}), \text{ where } p_{ij} = s_{x_j}(y_i) = (t^*)_{x_j}(y_i) = [(t_{y_j}(x_i))]^*,$$

while

$$\Psi(t) = M_t = (q_{ij}), \text{ where } q_{ij} = t_{y_j}(x_i).$$

Obviously,

$$p_{ij} = [t_{y_j}(x_i)]^* = [q_{ji}]^*.$$

Thus,

$$M_s = (p_{ij}) = ([q_{ji}]^*) = (q_{ij})^* = (M_t)^*,$$

$$\text{and we have } \Psi(t^*) = M_s = (M_t)^* = (\Psi(t))^*.$$

Q.E.D.

The following theorem, along with Lemmas 2.1 through 2.4, show how the embedding of the minimax algebra into the image algebra is accomplished.

Theorem 2.6. For a finite array X , with $|X| = m$,

$\{R_{\pm\infty}^X, \vee, \wedge; (R_{\pm\infty}^X)^X, \boxtimes, \vee, \boxtimes, \wedge; \boxtimes, \boxtimes\}$ is isomorphic to

$$\{R_{\pm\infty}^m, \vee, \wedge; M_{mm}, \times, \vee, \times', \wedge; \times, \times'\},$$

where M_{mm} is the set of all $m \times m$ matrices with entries in the bounded l -group $R_{\pm\infty}$.

Proof: By Lemma 2.1, ν preserves image-template multiplication, and by Lemma 2.2, ν

preserves the image-image pointwise maximum operation. By Lemmas 2.3 and 2.4,

for $X = Y = W$, Ψ preserves the operations of \boxtimes and \vee between templates. Let

$1 \in (R_{\pm\infty}^X)^X$ denote the identity template defined by

$$1_y(x) = \begin{cases} 0 & \text{if } y = x \\ -\infty & \text{otherwise} \end{cases}.$$

It is trivial to show that $\Psi(1) = e \in M_{mm}$, the identity matrix in M_{mm} .

We now show that the operations of \boxtimes and \wedge are also preserved under Ψ . It is not difficult to show that $\Psi(s \wedge t) = \Psi(s) \wedge \Psi(t)$. Let $r = s \wedge t$. Then

$\Psi(r) = M_r = (m_{ij}) = (r_{y_j}(x_i))$, where $r_{y_j}(x_i) = s_{y_j}(x_i) \wedge t_{y_j}(x_i)$. Thus,

$$\Psi(s) \wedge \Psi(t) = M_s \wedge M_t = (s_{y_j}(x_i)) \wedge (t_{y_j}(x_i)) = (s_{y_j}(x_i) \wedge t_{y_j}(x_i)) = (r_{y_j}(x_i)).$$

By definition, $s \boxtimes t = (t^* \boxtimes s^*)^*$, and, using Lemma 2.5 with $F = R$, Lemma 2.3, and property C_3 , we have

$$\begin{aligned} \Psi(s \boxtimes t) &= \Psi((t^* \boxtimes s^*)^*) = [\Psi(t^* \boxtimes s^*)]^* = (\Psi(t^*) \times \Psi(s^*))^* \\ &= (\Psi(s^*))^* \times' (\Psi(t^*))^* = \Psi(s) \times' \Psi(t). \end{aligned}$$

Thus, $\Psi(s \boxtimes t) = \Psi(s) \times' \Psi(t)$.

It is straightforward to see that ν is on-to-one and onto $R_{\pm\infty}^m$. To show that Ψ is one-one and onto M_{mm} , let $s, t \in (R_{\pm\infty}^X)^Y$ and suppose that $\Psi(s) = \Psi(t)$. Then

$$(\Psi(s))_{ij} = (M_s)_{ij} = s_{y_j}(x_i) = t_{y_j}(x_i) = (M_t)_{ij} = (\Psi(t))_{ij}, \text{ and, thus,}$$

$$s_{y_j}(x_i) = t_{y_j}(x_i) \text{ for all } j = 1, \dots, n, \text{ and for all } i = 1, \dots, m.$$

So Ψ is one-to-one as $s = t$. Let $M = (m_{ij}) \in M_{mn}$. Define $t \in (R_{\pm\infty}^X)^Y$ by

$$t_{y_j}(x_i) = m_{ij}. \text{ Then } \Psi(t) = M. \text{ Setting } m = n, \text{ we see that } \Psi \text{ is one-one and onto}$$

$$M_{mm}.$$

Q.E.D.

Thus, the minimax algebra with the bounded l-group $R_{\pm\infty}$ is embedded into image algebra, by the functions Ψ^{-1} and ν^{-1} . As the bounded l-group $R_{\pm\infty}^+$ is isomorphic to the bounded l-group $R_{\pm\infty}$ the minimax algebra with the bounded l-group $R_{\pm\infty}^+$ is also embedded into the image algebra. In this case, the matrix operation \times corresponds to the image algebra operation \odot . The isomorphism result is stated in Theorem 2.9.

Let X and Y be finite arrays as before. Let $R_{\pm\infty}^+$ have its usual meaning, $a \in (R_{\pm\infty}^+)^X$, M_{mn} denote the set of $m \times n$ matrices with values in $R_{\pm\infty}^+$ and let $(R_{\pm\infty}^+)^m = \{(x_1, x_2, \dots, x_m) : x_i \in R_{\pm\infty}^+\}$. Define $\nu : (R_{\pm\infty}^+)^X \rightarrow (R_{\pm\infty}^+)^m$ in the usual way by

$$\nu(a) = (a(x_1), \dots, a(x_m)).$$

Define $\Psi : ((R_{\pm\infty}^+)^X)^Y \rightarrow M_{mn}$ as before by

$$\Psi(t) = M_t = (p_{ij}), \text{ where } p_{ij} = t_{y_j}(x_i).$$

In the following lemmas, we assume that $|X| = m$, $|Y| = n$, and $|W| = l$. We claim the following, for $a, b \in (R_{\pm\infty}^+)^X$:

Lemma 2.7. $\nu(a \odot t) = \nu(a) \times \Psi(t)$, for $t \in ((R_{\pm\infty}^+)^X)^Y$.

Lemma 2.8. $\Psi(s \odot t) = \Psi(s) \times \Psi(t)$, for $s \in ((R_{\pm\infty}^+)^X)^W$, $t \in ((R_{\pm\infty}^+)^X)^Y$.

Proof to Lemma 2.7.

We must show that

$$(a \otimes t)(y_k) = (\nu(a) \times \Psi(t))_k.$$

We have

$$(a \otimes t)(y_k) = \bigvee_{x \in X} a(x) * t_{y_k}(x) = \bigvee_{i=1}^m a(x_i) * t_{y_k}(x_i).$$

$$\text{Also, } (\nu(a) \times \Psi(t))_k = \bigvee_{j=1}^m (\nu(a))_j * (\Psi(t))_{jk} = \bigvee_{j=1}^m a(x_j) * t_{y_k}(x_j).$$

Q.E.D.

Proof to Lemma 2.8.

Here, $s \in ((R_{\pm\infty}^+)^X)^W$ and $t \in ((R_{\pm\infty}^+)^W)^Y$ implies

$$s \otimes t = r \in ((R_{\pm\infty}^+)^X)^Y,$$

and

$$r_{y_j}(x_i) = \bigvee_{w \in W} t_{y_j}(w) * s_w(x_i) = \bigvee_{k=1}^l t_{y_j}(w_k) * s_{w_k}(x_i).$$

Now, let $\Psi(s) \times \Psi(t) = u \in M_{mn}$. We have

$$u_{ij} = \bigvee_{k=1}^l (\Psi(s))_{ik} * (\Psi(t))_{kj} = \bigvee_{k=1}^l s_{w_k}(x_i) * t_{y_j}(w_k) = \bigvee_{k=1}^l t_{y_j}(w_k) * s_{w_k}(x_i).$$

Q.E.D.

Theorem 2.9. For a finite array X , with $|X| = m$,

$\{((R_{\pm\infty}^+)^X)^X, \vee, \wedge; ((R_{\pm\infty}^+)^X)^X, \otimes, \vee, \otimes, \wedge; \otimes, \otimes\}$ is isomorphic to

$$\{(R_{\pm\infty}^+)^m, \vee, \wedge; M_{mm}, \times, \vee, \times', \wedge; \times, \times'\},$$

where M_{mm} is the set of all $m \times m$ matrices with entries in the bounded l -group $R_{\pm\infty}^+$.

Proof: By Lemma 2.7, ν preserves image-template multiplication, and by Lemma 2.2, ν

preserves the image-image pointwise maximum operation. By Lemmas 2.8 and 2.4.

for $X = Y = W$, Ψ preserves the operations of \boxtimes and \vee between templates. Let $1 \in ((R_{\pm\infty}^+)^X)^X$ denote the identity template defined by

$$1_y(x) = \begin{cases} 1 & \text{if } y = x \\ -\infty & \text{otherwise} \end{cases}$$

It is trivial to show that $\Psi(1) = e \in M_{mm}$, the identity matrix in M_{mm} over $R_{\pm\infty}^+$.

In Theorem 2.6, the proof that $\Psi(s \wedge t) = \Psi(s) \wedge \Psi(t)$ was not dependent on the value set $R_{\pm\infty}$ and hence is true also for templates $s, t \in ((R_{\pm\infty}^+)^X)^Y$. We now show that the operation of \otimes is also preserved under Ψ . By definition, $s \otimes t = \overline{(t \otimes \bar{s})}$, and, using Lemma 2.5 with $F = R^+$, Lemma 2.8, and property C_3 , we have

$$\begin{aligned} \Psi(s \otimes t) &= \Psi(\overline{(t \otimes \bar{s})}) = [\Psi(\bar{t} \otimes \bar{s})]^* = [\Psi(\bar{t}) \times \Psi(\bar{s})]^* \\ &= (\Psi(\bar{s}))^* \times' (\Psi(\bar{t}))^* = \Psi(s) \times' \Psi(t). \end{aligned}$$

Thus, $\Psi(s \otimes t) = \Psi(s) \times' \Psi(t)$.

We use the fact that Theorem 2.6 showed Ψ and ν are one-one and onto and also that $R_{\pm\infty}$ and $R_{\pm\infty}^+$ are isomorphic as bounded l-groups, and we are done.

Q.E.D.

We have shown that the minimax algebra with two different interpretations for the bounded l-group $F_{\pm\infty}$ with group F , namely $F = R$ and $F = R^+$, is embedded in the image algebra. Using the notation $R_{\pm\infty}^+$ instead of $R_{\pm\infty}^{\geq 0}$ allows the reader to regard the value sets $R_{\pm\infty}^+$ and $R_{\pm\infty}$ as basically the same (they are isomorphic as belts), without shifting gears from using 0 in one as the bottom element and $-\infty$ in the other. All minimax properties stated in Cuninghame-Green's book will be valid in the correct context of image algebra notation.

In using the minimax algebra results, we would like to point out that the the matrix-vector multiplication, multiplication of a matrix by a vector from the right, is used mostly throughout Cuninghame-Green's book. Left multiplication is mentioned at various places, and in fact, most left variants of the right multiplication results will hold. However, for the most part in our applications to image algebra, we will be using the right multiplication form in the development of our theory. The functions Ψ and ν map the image algebra expression $a \boxtimes t = b$ to the matrix algebra expression $\nu(a) \times \Psi(t) = \nu(b)$, the left multiplication form which we have omitted in our presentation of Cuninghame's material. The following diagram in Figure 5 explains how we will be taking advantage of the minimax algebra results.

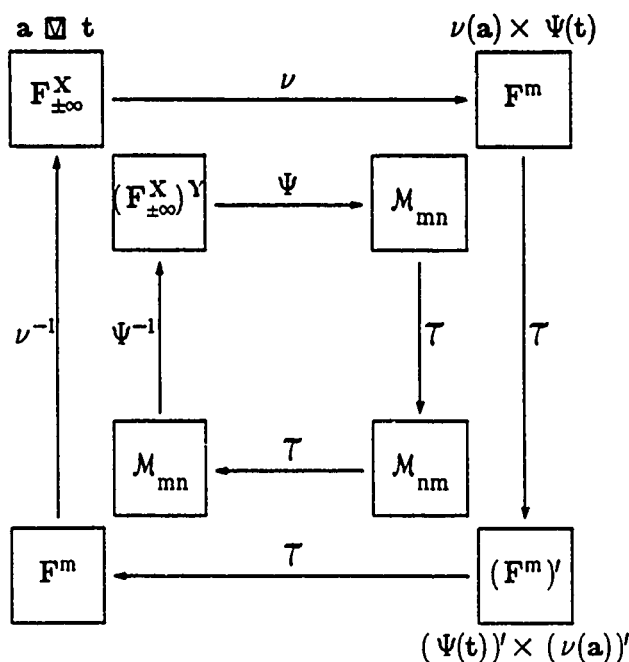


Figure 5. How the Transpose is used in Conjunction with the Isomorphism.

Let τ denote the function that takes a matrix to its transpose as well as the function that

takes a template to its transpose. Thus, $\mathcal{T}: \mathcal{M}_{mn} \rightarrow \mathcal{M}_{nm}$ is defined by

$$\mathcal{T}(\mathbf{a}) = \mathbf{a}',$$

the prime denoting as usual the transpose of a matrix, and $\mathcal{T}: (\mathbf{F}^X)^Y \rightarrow (\mathbf{F}^Y)^X$ is defined by

$$\mathcal{T}(\mathbf{t}) = \mathbf{t}'.$$

Obviously, $\Psi(\mathcal{T}(\mathbf{t})) = \mathcal{T}(\Psi(\mathbf{t}))$. In a clockwise manner, the functions ν and Ψ take the product $\nu(\mathbf{a} \boxtimes \mathbf{t})$ to $\nu(\mathbf{a}) \times \Psi(\mathbf{t})$, which is the matrix $\Psi(\mathbf{t})$ multiplied on the left by the row vector $\nu(\mathbf{a})$. Applying the transpose to $\nu(\mathbf{a}) \times \Psi(\mathbf{t})$, we get $\mathcal{T}[\nu(\mathbf{a}) \times \Psi(\mathbf{t})] = [\Psi(\mathbf{t})]' \times [\nu(\mathbf{a})]'$, which is the matrix $[\Psi(\mathbf{t})]' \in \mathcal{M}_{nm}$ multiplied on the right by the column vector $[\nu(\mathbf{a})]'$. We now use our minimax algebra theorems, where matrix-vector multiplication is the matrix multiplied on its right by a column vector. After getting the desired results, we continue on around the diagram clockwise, mapping back by the transpose \mathcal{T} again and then by ν^{-1} or Ψ^{-1} . Formally, if \mathbf{d} represents the column vector which is the result of applications of minimax algebra theorems to the initial column vector $(\Psi(\mathbf{t}))' \times (\nu(\mathbf{a}))'$, then $\nu^{-1}(\mathcal{T}(\mathbf{d}))$ will be an image. A similar situation holds for templates.

The minimax algebra results are stated in the usual matrix-vector multiplication order, and the isomorphisms Ψ and ν are used along with the transpose \mathcal{T} to apply the matrix results. When the word *isomorphism* is used in this context, it will mean the above functions Ψ and ν explicitly (not with the transpose) unless otherwise stated, with images as row vectors and templates as matrices with images t_y as columns.

PART II

MINIMAX APPLICATIONS TO IMAGE ALGEBRA AND IMAGE PROCESSING

The objective of the chapters in Part II is to show how the minimax algebra can be used to extend basic matrix algebraic results in such a way as to have applications in image processing. The tool that makes the minimax algebra useful in image processing is the isomorphism between the image algebra and the minimax algebra. Before the research presented in this dissertation was conducted, the relationship between the image algebra and the minimax algebra had not been established. The power of the isomorphism is that it makes all results in the minimax algebra applicable to solving image processing problems, just as linear algebra results are applicable to solving image processing problems. For example, template decomposition is presently a very active area of research. The problem of mapping transforms to some types of parallel architectures is equivalent to decomposing a transform t into a product of transforms $t = t^1 \boxtimes t^2 \boxtimes \cdots \boxtimes t^k$, where each factor t^i is directly implementable on the parallel architecture. Since decomposing templates is the same as decomposing matrices, matrix decomposition techniques can be applied to template decomposition problems. Thus far, there exist no decomposition techniques for matrices under the matrix operation \times as presented in section 1.2. Hence, the methods developed in Chapter 5 that decompose matrices are new results. They were developed mainly for solving the problem of mapping of transforms to particular parallel architectures, though they stand by themselves as a new theoretical result in the minimax algebra.

While some other areas of minimax algebra may seem to have no current applications to image processing, such as the eigenproblem, we present them in their image algebra form due to their interesting mathematical results.

CHAPTER 3

MAPPING OF MINMAX ALGEBRA PROPERTIES TO IMAGE ALGEBRA PROPERTIES

This chapter is devoted to describing algebraic properties of the substructures $\{(F^X)^Y, F^X, \boxtimes, \vee, \boxdot, \wedge\}$, where F is a subbelt of $R_{\pm\infty}$ or $R_{\pm\infty}^+$. During the investigation of the properties and before the discovery of the link to minimax algebra, many basic properties, such as the associativity of the \boxtimes operation, were proven within the context of the image algebra. Many theorems had excessive notational overhead, and often the proofs were laborious. Most of these same properties were found to have been stated and proven in context of the minimax algebra [38]. Using the matrix calculus makes some proofs less tedious, and in some cases makes them less cumbersome notationally. Thus, in order to place the presentation in a more elegant mathematical environment, we are omitting proofs that were done in the image algebra notation, and shall make use of the isomorphisms given in the previous chapter. Most of the theorems presented here are mapped into image algebra notation using the isomorphisms, and the proofs will be omitted. The results will be stated for both bounded l-groups, using the operations \boxtimes and \boxdot .

3.1. Basic Definitions and Properties

Unless otherwise stated, we shall assume that X , Y , and W are finite coordinate sets, with $|X| = m$, $|Y| = n$, $|W| = k$, with the pixel locations lexicographically ordered as in Chapter 2. The belt F with duality is a subbelt of either $R_{-\infty}$ or $R_{-\infty}^+$. The templates s and t will be F valued templates on appropriate domains, and a , b will be F valued images. For the appropriate subbelt F of $R_{\pm\infty}$ or $R_{\pm\infty}^+$ according to the operation \boxtimes or \boxdot , respectively, we have the following basic properties.

- (1) $((F^X)^Y, \vee)$ is an s -lattice and $((F^X)^Y, \vee)$ is a function space over (F, \vee, \times) ;
- (2) $\{(F^X)^X, \vee, \boxtimes\}$ is a belt; $\{(F^X)^X, \vee, \otimes\}$ is a belt;
- (3) $((F^X)^Y, \vee)$ is a left space over the belt $((F^X)^X, \vee, \boxtimes)$; $((F^X)^Y, \vee)$ is a left space over the belt $((F^X)^X, \vee, \otimes)$;
- (4) $(F^X)^Y$ is a right space over the belt F ;
- (5) We define *scalar multiplication* of a template $t \in (F^X)^Y$ by a scalar $\lambda \in F$ as multiplication by the one-point template $\lambda \in (F_{-\infty}^X)^X$ or $\lambda \in (F_{-\infty}^Y)^Y$, depending on whether the template λ multiplies from the left or from the right, respectively, (and adjoining $-\infty$ to F if necessary), as

$$t \boxtimes \lambda = \lambda \boxtimes t = s \in (F_{-\infty}^X)^Y, \text{ where } s_y(x) = t_y(x) + \lambda$$

and

$$t \otimes \lambda = \lambda \otimes t = s \in (F_{-\infty}^X)^Y, \text{ where } s_y(x) = t_y(x) * \lambda.$$

$$\text{Here, } \lambda_y(x) = \begin{cases} \lambda & \text{if } x = y \\ -\infty & \text{otherwise} \end{cases}$$

Next we state the distributive properties of \boxtimes and \otimes with respect to \vee .

- | | | |
|-----|---|---|
| (6) | $a \boxtimes (t \vee s) = (a \boxtimes t) \vee (a \boxtimes s)$ | $a \otimes (t \vee s) = (a \otimes t) \vee (a \otimes s)$ |
| | $a \boxtimes (t \boxtimes s) = (a \boxtimes t) \boxtimes s$ | $a \otimes (t \otimes s) = (a \otimes t) \otimes s$ |
| | $(a \vee b) \boxtimes t = (a \boxtimes t) \vee (b \boxtimes t)$ | $(a \vee b) \otimes t = (a \otimes t) \vee (b \otimes t)$ |
| | $(s \vee t) \boxtimes u = (s \boxtimes u) \vee (t \boxtimes u)$ | $(s \vee t) \otimes u = (s \otimes u) \vee (t \otimes u)$ |
| | $u \boxtimes (s \vee t) = (u \boxtimes s) \vee (u \boxtimes t)$ | $u \otimes (s \vee t) = (u \otimes s) \vee (u \otimes t)$ |
| | $s \boxtimes (t \boxtimes u) = (s \boxtimes t) \boxtimes u$ | $s \otimes (t \otimes u) = (s \otimes t) \otimes u.$ |

The dual to properties 1 through 6 also hold, as both the belts R and R^+ have duality.

- (7) $((F^X)^Y, \wedge)$ is an s -lattice and $((F^X)^Y, \wedge)$ is a function space over (F, \wedge, \times') ;
- (8) $\{(F^X)^X, \wedge, \boxtimes\}$ is a belt. $\{(F^X)^X, \wedge, \otimes\}$ is a belt.
- etc.

Now let F be a subbelt of R or R^+ , and $F_{\pm\infty}$ the bounded l-group with group F .

Corresponding to the identity matrix and the null matrix we have the *identity template*

$1 \in (F_{\pm\infty}^X)^X$, defined by

$$1_y(x) = \begin{cases} \phi & \text{if } x = y \\ -\infty & \text{otherwise} \end{cases}$$

and the *null template* $\Phi \in (F_{\pm\infty}^X)^Y$ defined by

$$\Phi_y(x) = -\infty, \text{ for all } y \in Y, x \in X.$$

For the belt R , $\phi = 0$, and for the belt R^+ , $\phi = 1$. Thus we have

$$a \boxtimes 1 = a, \quad t \boxtimes 1 = 1 \boxtimes t = t \quad \forall a \in R_{\pm\infty}^X, \forall t \in (R_{\pm\infty}^X)^X$$

For $\Phi \in (F^X)^X$,

$$t \vee \Phi = t, \quad t \boxtimes \Phi = \Phi \boxtimes t = \Phi, \quad a \boxtimes \Phi = \text{null image}, \quad \forall a \in R_{\pm\infty}^X, \forall t \in (R_{\pm\infty}^X)^X.$$

Similar properties hold for the operation \odot .

3.1.1. Homomorphisms

We now discuss homomorphisms in context of the image algebra. Let $|X| = m$. Since the s-lattice $\{F_{\pm\infty}^X, \vee\}$ is isomorphic (via ν) to the s-lattice $\{F_{\pm\infty}^m, \vee\}$, $\{F_{\pm\infty}^X, \vee\}$ is a space.

For $\lambda \in F^X$ the constant image, we have

$$a \vee \lambda = \lambda \vee a = b \in F_{\pm\infty}^X, \text{ where } b(x) = a(x) \vee \lambda,$$

and for the one-point template $\lambda \in (R_{\pm\infty}^X)^X$,

$$a \boxtimes \lambda = \lambda \boxtimes a = b \in F_{\pm\infty}^X, \text{ where } b(x) = a(x) + \lambda,$$

if $F = R$, and

$$a \odot \lambda = \lambda \odot a = b \in F_{\pm\infty}^X, \text{ where } b(x) = a(x) * \lambda,$$

if $F = R^+$. Let $F \in \{R_{\pm\infty}, R_{\pm\infty}^+\}$. Since $\{F^Y, \vee\}$ is an s-lattice, an s-lattice homomorphism

from F^X to F^Y is a function $f: F^X \rightarrow F^Y$ satisfying

$$f(a \vee b) = f(a) \vee f(b).$$

A right linear homomorphism $g: F^X \rightarrow F^Y$ is an s -lattice homomorphism satisfying

$$g(a \boxtimes \lambda) = g(a) \boxtimes \lambda.$$

Thus, the set of all right linear homomorphisms from F^X to F^Y is denoted by

$$\text{Hom}_F(F^X, F^Y) = \{g: F^X \rightarrow F^Y, \text{ and } g \text{ satisfies } g(a \vee b) = g(a) \vee g(b), g(a \boxtimes \lambda) = g(a) \boxtimes \lambda\},$$

or if F is R^+ , then

$$\text{Hom}_F(F^X, F^Y) = \{g: F^X \rightarrow F^Y, \text{ and } g \text{ satisfies } g(a \vee b) = g(a) \vee g(b), g(a \otimes \lambda) = g(a) \otimes \lambda\}.$$

3.1.2. Classification of Homomorphisms in the Image Algebra

Right linear transformations can be characterized entirely in terms of template transformations, and we give necessary and sufficient conditions for $(F^X)^Y$ to be isomorphic to $\text{Hom}_F(F^X, F^Y)$.

Theorem 3.1. *Let F be a belt with identity and null element. Then for all non-empty finite coordinate sets X, Y , $(F^X)^Y$ is isomorphic to $\text{Hom}_F(F^X, F^Y)$.*

Corollary 3.2. *Let F be a belt, and let $X \neq \emptyset$ be a finite coordinate set with $|X| \geq 1$.*

Then a necessary and sufficient condition that $(F^X)^Y$ be isomorphic to $\text{Hom}_F(F^X, F^Y)$, for all non-empty finite coordinate sets Y , is that F have an identity element ϕ with respect to \times and a null element θ with respect to \vee .

We call a template $t \in (F^X)^Y$ used with the operation $\vee, \boxtimes, \boxdot, \otimes$, or \odot a *lattice transform*. We will present an example of a transformation which is not right linear in section 6.1.

3.1.3. Inequalities

Some useful inequalities are stated in the next theorem.

Theorem 3.3. *Let F be a subbelt of $R_{\pm\infty}$ or $R_{\pm\infty}^+$. Then the following inequalities hold for images and templates with the appropriate domains, having values in F .*

- (i) $a \vee (b \wedge c) \leq (a \vee b) \wedge (a \vee c)$
- (ii) $a \wedge (b \vee c) \geq (a \wedge b) \vee (a \wedge c)$
- (iii) $(a \wedge b) \boxtimes t \leq (a \boxtimes t) \wedge (b \boxtimes t) \quad (a \wedge b) \oslash t \leq (a \oslash t) \wedge (b \oslash t)$
- (iv) $a \boxtimes (t \wedge s) \leq (a \boxtimes t) \wedge (a \boxtimes s) \quad a \oslash (t \wedge s) \leq (a \oslash t) \wedge (a \oslash s)$
- (v) $(a \vee b) \boxtimes t \geq (a \boxtimes t) \vee (b \boxtimes t) \quad (a \vee b) \oslash t \geq (a \oslash t) \vee (b \oslash t)$
- (vi) $a \boxtimes (t \vee s) \geq (a \boxtimes t) \vee (a \boxtimes s) \quad a \oslash (t \vee s) \geq (a \oslash t) \vee (a \oslash s)$

- (i) $s \vee (t \wedge r) \leq (s \vee t) \wedge (s \vee r)$
- (ii) $s \wedge (t \vee r) \geq (s \wedge t) \vee (s \wedge r)$
- (iii) $t \boxtimes (s \wedge r) \leq (t \boxtimes s) \wedge (t \boxtimes r) \quad t \oslash (s \wedge r) \leq (t \oslash s) \wedge (t \oslash r)$
- (iv) $(s \wedge r) \boxtimes t \leq (s \boxtimes t) \wedge (r \boxtimes t) \quad (s \wedge r) \oslash t \leq (s \oslash t) \wedge (r \oslash t)$
- (v) $t \boxtimes (s \vee r) \geq (t \boxtimes s) \vee (t \boxtimes r) \quad t \oslash (s \vee r) \geq (t \oslash s) \vee (t \oslash r)$
- (vi) $(s \vee r) \boxtimes t \geq (s \boxtimes t) \vee (r \boxtimes t) \quad (s \vee r) \oslash t \geq (s \oslash t) \vee (r \oslash t)$

$$a \boxtimes (s \boxtimes r) \leq (a \boxtimes s) \boxtimes r \text{ and } a \boxtimes (s \boxtimes r) \geq (a \boxtimes s) \boxtimes r$$

$$t \boxtimes (s \boxtimes r) \leq (t \boxtimes s) \boxtimes r \text{ and } t \boxtimes (s \boxtimes r) \geq (t \boxtimes s) \boxtimes r$$

and

$$a \oslash (s \oslash r) \leq (a \oslash s) \oslash r \text{ and } a \oslash (s \oslash r) \geq (a \oslash s) \oslash r.$$

$$t \oslash (s \oslash r) \leq (t \oslash s) \oslash r \text{ and } t \oslash (s \oslash r) \geq (t \oslash s) \oslash r.$$

We remark that the above properties corresponding to the *forward* multiplications of an image by a template as defined in Chapter 1 are also valid, namely,

$$t \boxtimes (a \wedge b) \leq (t \boxtimes a) \wedge (t \boxtimes b), \text{ etc.}$$

3.1.4. Conjugacy

The notion of conjugacy as discussed in section 1.2 extends to templates as well. Suppose that F and F^* are conjugate. Then for $t \in (F^X)^Y$, $t^* \in ((F^*)^Y)^X$ is defined by

$$t_x^*(y) \equiv (t_y(x))^*.$$

The conjugate of $t \in (R_{\pm\infty}^X)^Y$ is the additive dual t^* , and the conjugate of $t \in ((R_{\pm\infty}^+)^X)^Y$ is the multiplicative dual \bar{t} , both of which are defined in section 1.1.

Let P be any set of F valued templates from Y to X , with F and F^* as conjugate systems. Define P^* by

$$P^* \equiv \{t^* : t \in P\}.$$

Here, the star symbol $*$ denotes the dual template for either value set $R_{\pm\infty}$ or $R_{\pm\infty}^+$. Note that $P^* \subset ((F^*)^Y)^X$. We have

Theorem 3.4. *Let (F, \vee) and (F^*, \wedge) be conjugate. Then $((F^X)^Y, \vee, \boxtimes)$ and $((F^*)^Y)^X, \wedge, \boxtimes)$ are conjugate, where F is a sub-bounded l-group of $R_{\pm\infty}$ and $((F^X)^Y, \vee, \odot)$ and $((F^*)^Y)^X, \wedge, \odot)$ are conjugate, where F is a sub-bounded l-group of $R_{\pm\infty}^+$ for any non-empty finite coordinate sets X, Y . In all cases the conjugate of a given template t is the dual template t^* or \bar{t} of the respective bounded l-group as defined in Chapter 1.*

Proposition 3.5. *If $(F, \vee, \times, \wedge, \times')$ is a self-conjugate belt, then $((F^*)^X)^Y = (F^X)^Y$ for all non-empty finite coordinate sets X, Y . Also, $((R_{\pm\infty}^X)^X, \vee, \boxtimes, \wedge, \boxtimes)$ is a self-conjugate belt, and $((R_{\pm\infty}^+)^X)^X, \vee, \odot, \wedge, \odot)$ is a self-conjugate belt.*

An example. In this section we give an application to a scheduling problem, showing the use of the conjugate of a template. In particular, this example provides a physical interpretation of the conjugate of a template.

Suppose we have n tasks, or activities, or subroutines, labelled $1, \dots, n$. Let $a(x_i)$ denote the starting time of task i , and assume without loss of generality that task 1 is the starting activity, task n is the finishing activity, and that tasks 2 through $n-1$ are intermediate activities. Suppose we are given the time of the starting activity, and we wish to know the soonest time at which each subsequent activity can be started. In particular, what is the earliest time that task n can start, or, what is the earliest expected time of completion of the collection of tasks?

The relation of the tasks to one another can be described by a partial order \mathcal{R} on the set of tasks $\{1, \dots, n\}$:

$j \mathcal{R} i$ if and only if task j is to be completed before task i can start.

Let d_{ij} denote the minimum amount of time by which the start of activity j must precede the start of activity i . That is, d_{ij} is the duration time of activity j , or the processing time of task j , which must pass before activity i can start. Define $w \in (\mathbb{R}_{-\infty}^X)^X$ by

$$w_{x_i}(x_j) = \begin{cases} d_{ij} & \text{if } j \mathcal{R} i \\ -\infty & \text{otherwise} \end{cases}.$$

There is an obvious relationship between the weighted digraph associated with the partial order relation \mathcal{R} and the template w . For example, suppose we have 5 tasks or activities, or subroutines of a program, which have the following relation or partial order:

$$(1,2) \ (1,3) \ (2,4) \ (2,5) \ (3,4) \ (3,5) \ (4,5)$$

Here, activity 1 is the start activity, activity 5 is the end activity, and tasks 2,3,4 are inter-

mediate tasks or subroutines. Suppose the duration times d_{ij} of the activities are:

$$d_{21} = 1 \quad d_{31} = 6 \quad d_{42} = 2$$

$$d_{43} = 1 \quad d_{52} = 1 \quad d_{53} = 3 \quad d_{54} = 3$$

and $d_{ii} = 0$ for each $i = 1, \dots, 5$. This is consistent with a meaningful physical interpretation of the definition of duration time for a task.

The corresponding weighted digraph is given in Figure 6.

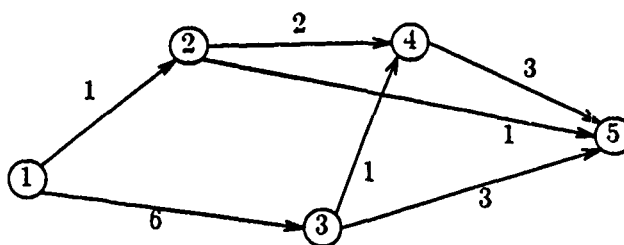


Figure 6. A Scheduling Network.

The nodes represent the activities, and the duration times are given as numbers on the directed edges linking the nodes.

In determining $a(x_4)$ for example, note that $a(x_4)$ must satisfy

$$a(x_4) = \max\{d_{42} + a(x_2), d_{43} + a(x_3), d_{44} + a(x_4)\}$$

or equivalently

$$a(x_4) = \max_{1 \leq j \leq 5} \{w_{x_4}(x_j) + a(x_j)\}.$$

This last equality follows from the fact that $w_{x_i}(x_j) = -\infty$ if j is not related to i . In the general setting, we must solve, for each $i = 1, \dots, n$:

$$a(x_i) = \max_{1 \leq j \leq n} \{w_{x_i}(x_j) + a(x_j)\}$$

or, writing the problem as an image algebra expression, we must solve for a in

$$a \boxminus w = a. \quad (3-1)$$

Here, a is an image on X where $|X| = n$.

An analysis of a network in this manner is called *backward recursion analysis*.

Under *forward recursion*, suppose we have n tasks with duration times f_{ij} , where f_{ij} is the minimum amount of time by which the start of activity i must precede the start of activity j , if the activities are so related. Otherwise, let f_{ij} have value $-\infty$. Define $w \in (R_{\pm\infty}^X)^X$ by

$$w_{x_i}(x_j) = \begin{cases} f_{ij} & \text{if } j \mathcal{R} i \\ -\infty & \text{otherwise} \end{cases}$$

As before, $f_{ii} = 0$ gives a consistent physical interpretation.

Let τ be the planned completion date of the project, which is given, and define $a(x_i)$ to be the latest allowable starting time for activity i . We wish to determine $a(x_1), \dots, a(x_{n-1})$ such that $a(x_n) = \tau$. Thus, we desire to solve for a in

$$a(x_i) = \min_{j=1, \dots, n} (-w_{x_i}(x_j) + a(x_j))$$

for $i = 1, \dots, n$. For example, for 5 nodes, suppose we have the following relations:

$$(1,2) \ (1,3) \ (2,4) \ (2,5) \ (3,4) \ (3,5) \ (4,5).$$

Here we write (i,j) if task i must precede task j . Suppose the times f_{ij} of the activities are:

$$f_{12} = 1 \quad f_{13} = 6 \quad f_{24} = 2$$

$$f_{34} = 1 \quad f_{25} = 1 \quad f_{35} = 3 \quad f_{45} = 3$$

Suppose we would like to find $a(x_4)$, say, satisfying

$$a(x_4) = \min_{j=1,\dots,5} (-w_{x_4}(x_j) + a(x_j)).$$

The value $-w_{x_4}(x_5) + a(x_5)$ is the latest allowable time to start task 5 minus the minimum amount of time activity 4 must precede activity 5, and the time to start task 4 must be at least as small as this number. Thus, the time to start task 5 must be at least as small as $-1 + a(x_5)$. The value $a(x_4) = \min \{-w_{x_4}(x_5) + a(x_5)\} = -1 + \tau$. (All other values $-w_{x_4}(x_j) + a(x_j) = +\infty$ as $-w_{x_4}(x_j) = +\infty$ for $j \neq 5$.) Since τ is given, this quantity can be explicitly determined. The remaining equations can be solved similarly.

If we define $u \in (R_{+\infty}^X)^X$ by

$$u_{x_i}(x_j) = \begin{cases} -w_{x_i}(x_j) & \text{if } j \mathcal{R} i \\ +\infty & \text{otherwise} \end{cases}$$

then it is obvious that in general we must solve for a the following:

$$a \boxtimes u = a. \quad (3-2)$$

It is clear that the template u in equation (3-2) is the conjugate of the template w in Equation (3-1). That is,

$$u = w^*.$$

We can say that the templates w and w^* define the structure of the network as we analyze it backward or forward in time, respectively.

3.1.5. Alternating tt^* and $\bar{t}t$ Products

This section discusses the concept of an *alternating tt^* or $\bar{t}t$ product* of a template t and its conjugate under the operation \boxtimes or \boxdot , respectively. We shall state the results only for the sub-bounded l-groups of $R_{\pm\infty}$ and the operations \boxtimes and \boxdot , with the understanding that unless otherwise stated, an arbitrary sub-bounded l-group of $R_{\pm\infty}^+$ and the operations \boxdot and \boxtimes may be substituted in the appropriate places.

Theorem 3.8. Let $F_{\pm\infty}$ be a sub-bounded l -group of $R_{\pm\infty}$, where F denotes the group of the bounded l -group $F_{\pm\infty}$, and $t \in (F_{\pm\infty}^X)^Y$. Then we have

$$t \boxtimes (t^* \boxtimes t) = t \boxtimes (t^* \boxtimes t) = (t \boxtimes t^*) \boxtimes t = (t \boxtimes t^*) \boxtimes t = t.$$

Similarly,

$$t^* \boxtimes (t \boxtimes t^*) = t^* \boxtimes (t \boxtimes t^*) = (t^* \boxtimes t) \boxtimes t^* = (t^* \boxtimes t) \boxtimes t^* = t^*.$$

We now define an *alternating tt^* product*. Write a word consisting of the letters t and t^* , in an alternating sequence. A single letter t or t^* is allowed. If we have $k > 1$ letters, now insert $k-1$ symbols of \boxtimes and \boxtimes , in an alternating manner. For example, the following sequences are allowed:

$$t^* \boxtimes t$$

$$t \boxtimes t^* \boxtimes t$$

$$t^* \boxtimes t \boxtimes t^* \boxtimes t \boxtimes t^* \boxtimes t.$$

Now insert brackets in an arbitrary way so that the resulting expression is not ambiguous.

For example,

$$t^* \boxtimes t$$

$$t \boxtimes (t^* \boxtimes t)$$

$$(t^* \boxtimes ((t \boxtimes t^*) \boxtimes t)) \boxtimes (t^* \boxtimes t).$$

Any algebraic expression so constructed is called an *alternating tt^* product*.

Suppose an alternating tt^* product an odd number of letters t and/or t^* . Then we say it is of *type t* if it begins and ends with an t , and that it is of *type t^** if it begins and ends with an t^* . If it has an even number of letters we say that it is of *type*

$$t \boxtimes t^* \text{ or } t \boxtimes t^* \text{ or } t^* \boxtimes t$$

exactly according to the first two letters with its separating operator, regardless of how the

brackets lie in the entire expression. As an example:

$$\begin{aligned}
 t^* \boxtimes t & \text{ is of type } t^* \boxtimes t \\
 t \boxtimes (t^* \boxtimes t) & \text{ is of type } t \\
 (t^* \boxtimes ((t \boxtimes t^*) \boxtimes t)) \boxtimes (t^* \boxtimes t) & \text{ is of type } t^* \boxtimes t.
 \end{aligned}$$

Theorem 3.7. *Let $F_{\pm\infty}$ be a sub-bounded l-group of $R_{\pm\infty}$ and t an arbitrary template in $(F_{\pm\infty}^X)^Y$. Then every alternating tt^* product P is well-defined, and if P is of type Q , then $P = Q$.*

If a product P has more than 1 letter, then we define $P(z)$ to be the formal product obtained when the last (rightmost) letter, t or t^* (or \bar{t}), is replaced by z , where z is a F valued template on the appropriate coordinate sets X and Y .

Theorem 3.8. *Let $F_{\pm\infty}$ be a sub-bounded l-group of $R_{\pm\infty}$ and t, z arbitrary templates over F . If P is an alternating tt^* product containing four letters and P is of type Q , then*

$$P(z) = Q(z).$$

3.2. Systems of Equations

We now discuss the problem of finding solutions to the problem:

$$\text{Given } t \in (R_{\pm\infty}^X)^Y \text{ and } b \in R_{\pm\infty}^Y, \text{ find } a \in R_{\pm\infty}^X \text{ such that } a \boxtimes t = b. \quad (3-3)$$

Similarly, we also wish to solve:

$$\text{Given } t \in ((R_{\pm\infty}^+)^X)^Y \text{ and } b \in (R_{\pm\infty}^+)^Y, \text{ find } a \in (R_{\pm\infty}^+)^X \text{ such that } a \otimes t = b.$$

Here, $|X| = m, |Y| = n$.

3.2.1. F-asticity and /-solutions

If F is a bounded l-group and $x, y \in F$, we say that the products $x \times y$ and $x \times' y$ are /-undefined if one of x, y is $-\infty$ and the other is $+\infty$. We say that a template product is /-undefined if the evaluation of $t_y(x)$ requires the formation of a /-undefined product of elements of the bounded l-group $F_{\pm\infty}$. Otherwise, we say that a template product is /-defined or /-exists. Some mathematical models require solutions which avoid the formation of /-undefined products, as in practical cases these often correspond to unrelated activities. We state these results for both bounded l-groups where appropriate, with the \odot results in parentheses. As usual, the sub-bounded l-group $F_{\pm\infty}$ is dependent on which operation, \boxtimes or \odot is used.

Lemma 3.9. *Let $F_{\pm\infty}$ be a subbelt of $R_{\pm\infty}(R_{\pm\infty}^+)$. Let X and Y be non-empty, finite arrays, and $t \in (F_{\pm\infty}^X)^Y$. Then the set of all images $a \in F_{\pm\infty}^X$ such that $a \boxtimes t (a \odot t)$ is /-defined is a sub-s-lattice of $F_{\pm\infty}^X$. Hence the set of solutions a of statement (3-3) such that $a \boxtimes t (a \odot t)$ /-exists is either empty or is a sub-s-lattice of $F_{\pm\infty}^X$.*

Lemma 3.10. *Let X, Y , and W be non-empty, finite arrays, and $t \in (F_{\pm\infty}^W)^Y$. Then the set of templates $s \in (F_{\pm\infty}^X)^W$, such that $s \boxtimes t (s \odot t)$ is /-defined is a sub-s-lattice of $(F_{\pm\infty}^X)^Y$.*

Any solution a of statement (3-3) such that $a \boxtimes t (a \odot t)$ /-exists is called a /-solution of (3-3).

Theorem 3.11. *Let $F_{\pm\infty}$ be a sub-bounded l-group of $R_{\pm\infty}(R_{\pm\infty}^+)$. Then (3-3) has at least one solution if and only if $a = b \boxtimes t^* (a = b \odot \bar{t})$ is a solution. In this case, $a = b \boxtimes t^* (a = b \odot \bar{t})$ is the greatest solution.*

Recall from probability theory that a row-stochastic matrix is a non-negative matrix in which the sum of the elements in each row is equal to 1. We will make analogous definitions, where the operation $+$ is replaced by the operation \vee , and the unity element is $-\infty$.

Let $P \subset F_{\pm\infty}$, where $F_{\pm\infty}$ is an arbitrary sub-bounded l-group of $R_{\pm\infty}(R_{\pm\infty}^+)$. A template $t \in (F_{\pm\infty}^X)^Y$ is called *row-P-astic* if $\bigvee_{j=1}^n t_{y_i}(x_j) \in P$ for all $i = 1, \dots, n$ and *column-P-astic* if $\bigvee_{i=1}^n t'_{x_j}(y_i) \in P$ for all $j = 1, \dots, m$. The template t is called *doubly-P-astic* if t is both row- and column-P-astic. Note that if t is column-P-astic, then t' is row-P-astic.

Theorem 3.12. *Let $F_{\pm\infty}$ be a sub-bounded l-group of $R_{\pm\infty}(R_{\pm\infty}^+)$ and $t \in (F_{\pm\infty}^X)^Y$, $b \in F_{\pm\infty}^Y$ such that (3-3) is soluble. Then $a = b \boxtimes t^*$ ($a = b \boxtimes \bar{t}$) /-exists and is a /-solution of (3-3), if and only if one of the following cases is satisfied:*

- (i) $t \in (F_{\pm\infty}^X)^Y$, and $b = +\infty$, the constant image with $+\infty$ everywhere.
- (ii) $t \in (F_{\pm\infty}^X)^Y$, and $b = -\infty$.
- (iii) $t \in (F_{\pm\infty}^X)^Y$ is doubly F -astic, and $b \in F^X$.

Moreover, every solution of (3-3) is then a /-solution, and $b \boxtimes t^*$ ($b \boxtimes \bar{t}$) is equal to $+\infty$, $-\infty$, or is finite, respectively according as case (i), (ii), or (iii) holds.

In the following theorem, we state the dual and left-right generalizations of Theorems 3.11 and 3.12.

Corollary 3.13. *Let $F_{\pm\infty}$ be a sub-bounded l-group of $R_{\pm\infty}(R_{\pm\infty}^+)$, and let $t \in (F_{\pm\infty}^X)^Y$, $b \in F_{\pm\infty}^Y$. Then for all combinations of c, q , and δ given in Table 1, the following statement is true:*

The image algebra equation c has at least one solution if and only if the product d is a solution; and the product d is then the δ solution. Furthermore, if the product d is /-defined, and equation c is /-defined when $a = d$, then equation c is /-defined when a is any solution of equation c . If $F_{\pm\infty}$ is a sub-bounded l-group of $R_{\pm\infty}^+$ then

the results in Table 1 hold for \odot replacing \boxtimes everywhere and \bar{t} replacing t^* everywhere.

Table 1.

c	d	δ
$a \boxtimes t = b$	$b \boxtimes t^*$	greatest
$a \boxtimes t^* = b$	$b \boxtimes t$	greatest
$a \boxtimes t = b$	$b \boxtimes t^*$	least
$a \boxtimes t^* = b$	$b \boxtimes t$	least
$t \boxtimes a = b$	$t^* \boxtimes b$	greatest
$t^* \boxtimes a = b$	$t \boxtimes b$	greatest
$t \boxtimes a = b$	$t^* \boxtimes b$	least
$t^* \boxtimes a = b$	$t \boxtimes b$	least

If d is a solution to c in Table 1, then d is called a *principal solution*.

We can also restate the last three theorems as a solubility criterion.

Problem (3-3) is soluble if and only if $(b \boxtimes t^*) \boxtimes t = b [(b \boxtimes \bar{t}) \odot t = b]$; and every solution is a \wedge -solution if $(b \boxtimes t^*) \boxtimes t [(b \boxtimes \bar{t}) \odot t = b] \wedge$ -exists.

Note that Theorem 3.12 identifies the cases in which (3-3) has a \wedge -defined \wedge -solution.

All solutions are then \wedge -solutions. The next question to ask is: can we find all solutions? We now focus on the following problem.

Given that F is $R_{\pm\infty}(R_{\pm\infty}^+)$ and that $(b \boxtimes t^*) \boxtimes t) [(b \otimes \bar{t}) \otimes t)]$ /-exists and equals b , find all solutions of (3-3). (3-4)

For cases (i) and (ii) of Theorem 3.12, we note that t is finite. The next proposition gives solutions for these two cases.

Proposition 3.14. *Let $F_{\pm\infty}$ be a sub-bounded l-group of $R_{\pm\infty}(R_{\pm\infty}^+)$. If $b = -\infty$ (the constant image), then Problem (3-4) has b as its unique solution. If $b = +\infty$, then Problem (3-4) has as its solutions exactly those images of $F_{\pm\infty}^X$ which have at least one pixel value equal to $+\infty$.*

To determine solutions to case (iii), we need to consider the particular case that $F_{\pm\infty}$ is the 3-element bounded l-group F_3 . Here b is finite with all elements having value ϕ .

Lemma 3.15. *Let $F_{\pm\infty}$ be the 3-element bounded l-group F_3 . Let t be doubly F -astic and b be finite. Then (3-3) is soluble, having as principal /-solution $a = 1$ where $1(x) = \phi$ for all x . Hence, no solution to (3-3) contains $+\infty$ for any pixel value, and all solutions are /-solutions.*

3.2.2. All Solutions to $a \boxtimes t = b$ and $a \otimes t = b$

We now give some criterions for finding all solutions to problem (3-3) for the case where the template t is doubly F -astic and b finite. We discuss the general case where F is the belt R or R^+ .

If a template $t \in (F_{\pm\infty}^X)^X$ has form

$$t_{x_i}(x_i) = \alpha_i, \text{ and } t_{x_i}(x_j) = -\infty, j \neq i,$$

we write $t = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_m)$.

For $b \in F$ finite, define the template $d \in (F_{-\infty}^X)^X$ by $d =$

$$\text{diag}([b(x_1)]^*, [b(x_2)]^*, \dots, [b(x_m)]^*).$$

Since b is finite, so is $d_{x_i}(x_i)$, and $d_{x_i}(x_i) = -b(x_i)$ (or $1/b(x_i)$) $\forall i = 1, \dots, m$. Thus, solving (3-3) is equivalent to solving

$$a \boxtimes s = 1, \quad (3-5)$$

or

$$a \odot s = 1,$$

where $s = d \boxtimes t$ ($s = d \odot t$) $\in (F_{\pm\infty}^X)^Y$, and $1 = \phi$, the constant image. Note that

$s_{y_k}(x_j) = t_{y_k}(x_j) - b(x_j)$ ($s_{y_k}(x_j) = t_{y_k}(x_j) * 1/b(x_j)$). Now, for each image $s'_{x_j} \in F_{\pm\infty}^X$, let $W'_j = \{(x_j, y_i) : s'_{x_j}(y_i) = \bigvee_{k=1}^m s'_{x_j}(y_k)\}$. Note that $W'_j \subset X \times Y$ for every j . The elements $s'_{x_j}(y_i)$ corresponding to $(x_j, y_i) \in W'_j$ are called *marked values* of W'_j . Notice that every image s'_{x_j} will have at least one marked value, as d , t and s are doubly F -astic. Our next theorem gives conditions where there is no solution.

Lemma 3.16. *Let $F_{\pm\infty}$ be a bounded l -group, $t \in (F_{-\infty}^X)^Y$ where t is doubly F -astic, and $b \in F^Y$. Define $s \in (F_{-\infty}^X)^Y$ by*

$$s = d \boxtimes t \quad (\text{or} \quad s = d \odot t)$$

depending on whether the group F is R or R^+ , respectively, and d is as above. Suppose there exists i such that for no j is $s_{y_i}(x_j)$ a marked value. That is, suppose there exists $y_i \in Y$ such that $s_{y_i}(x_j)$ is not a marked value for any j . Then there does not exist $a \in F_{\pm\infty}^X$ such that $a \boxtimes t = b$ ($a \odot t = b$).

There now remains the case in which for every i , there is at least one j such that $s_{y_i}(x_j)$ is a marked value. We transform the question into a boolean problem, where it can be

shown that the following procedure will give a set of solutions to equation (3-5) [39].

Step 1. For the bounded l-group $F_{\pm\infty} = F_3$, define $g \in (F_{\pm\infty}^X)^Y$ by

$$g_{y_i}(x_j) = \begin{cases} \phi & \text{if } s'_{x_j}(y_i) \text{ is marked} \\ -\infty & \text{otherwise} \end{cases}.$$

Letting $f \in F_{\pm\infty}^X$, now solve the boolean system

$$f \boxtimes g = 1 \quad (\text{or} \quad f \odot g = \phi). \quad (3-6)$$

As in the case for matrices [39], each solution to equation (3-6) consists of an assignment of one of the values $-\infty$ or ϕ to each $f(x_j)$.

Let $f = (f(x_1), \dots, f(x_m))$ be a solution to equation (3-6).

Step 2. For each $j = 1, \dots, m$: if $f(x_j) = \phi$ then set $a(x_j)$ to be the value $-\left(\bigvee s'_{x_j}\right)$ $(1/\left(\bigvee s'_{x_j}\right))$. If $f(x_j) = -\infty$ then $a(x_j)$ is given an arbitrary value such that $a(x_j) < -\left(\bigvee s'_{x_j}\right) (1/\left(\bigvee s'_{x_j}\right))$.

For the boolean case, we have

Proposition 3.17. *The solutions of equation (3-6) are exactly the assignments of the values ϕ or $-\infty$ to the variables $f(x_j)$ such that for every $i = 1, \dots, m$ there holds $f(x_j) = \phi$ for at least one j such that $s_{y_i}(x_j)$ is a marked value.*

Theorem 3.18. *Let $F_{\pm\infty}$ be a bounded l-group. Then the above two step procedure yields all solutions to equation (3-7) without repetition.*

3.2.3. Existence and Uniqueness

This section discusses some existence and uniqueness theorems concerning solutions to Problem (3-3).

Theorem 3.19. Let $F_{\pm\infty}$ be a bounded l-group, and let $t \in (F_{-\infty}^X)^Y$ be doubly F-astic and $b \in F^Y$ be finite. Then a necessary and sufficient condition that the equation $a \boxtimes t = b$ ($a \odot t = b$) shall have at least one solution is that for all $x_i \in X$, there exists at least one j such that for the template $s = d \boxtimes t$ ($s = d \odot t$), where d is as defined as above,

$$s_{y_i}(x_j) \text{ is a marked value.}$$

We remark that the solution $a(x_i) = - \left(\bigvee s'_{x_j} \right) (1 / (\bigvee s'_{x_j}))$ gives exactly the principal solution.

This is equivalent to

Theorem 3.20. Let $F_{\pm\infty}$ be a bounded l-group, let $t \in (F_{-\infty}^X)^Y$ be doubly F-astic, and let $b \in F^Y$ be finite. Then a necessary and sufficient condition that the equation $a \boxtimes t = b$ ($a \odot t = b$) shall have exactly one solution is that for all $x_i \in X$, there exists at least one j such that

$$s_{y_i}(x_j) \text{ is a marked value,}$$

and for each $j = 1, \dots, n$, there exists an i , $1 \leq i \leq m$ such that $|W'_i| = 1$.

Define a template $t \in (F_{\pm\infty}^X)^Y$ to be strictly doubly ϕ -astic if it satisfies the following two conditions.

- (i) $t_{y_i}(x_j) \leq \phi$, $i, j = 1, \dots, n$
- (ii) for each $i = 1, \dots, n$, there exists a unique index $j \in \{1, 2, \dots, n\}$ such that $t_{y_i}(x_j)$ has value ϕ .

If $t \in (F_{\pm\infty}^X)^Y$, $|X| = m$, $|Y| = n$, then we say that t contains a template

$s \in (F_{\pm\infty}^{W_2})^{W_1}$ if the matrix $\Psi^{-1}(t)$ contains the matrix $\Psi^{-1}(s)$ of size $h \times k$, where $|W_2| = h$,

$|W_1| = k$, and both $h, k \leq \min(m, n)$. We say that a template $t \in (F_{\pm\infty}^X)^Y$ contains an image

$a \in F_{\pm\infty}^X$ if $a = t_y$ for some $y \in Y$.

Theorem 3.21. Let $F_{\pm\infty}$ be a bounded l-group, let $t \in (F_{\pm\infty}^X)^Y$ be doubly F-astic, and let $b \in F^Y$ be finite. Then a necessary and sufficient condition that the equation $a \boxtimes t = b$ ($a \otimes t = b$) shall have exactly one solution is that we can find k finite elements $\alpha_1, \dots, \alpha_k$ such that the template d defined by

$$d_{y_i}(x_j) = -b(y_i) + t_{y_i}(x_j) + \alpha_j \quad (\text{or } d_{y_i}(x_j) = b(y_i)^{-1} * t_{y_i}(x_j) * \alpha_j)$$

is doubly ϕ -astic and that d contains a strictly doubly ϕ -astic template $s \in (F_{-\infty}^W)^W$, $|W| = k$.

3.2.4. A Linear Programming Criterion

Since one of our interests is the case where the bounded l-group is the $R_{\pm\infty}$ we now show that the problem can be stated as a linear programming problem for this bounded l-group.

Theorem 3.22. Let $t \in (R_{\pm\infty}^X)^Y$ be doubly F-astic and $b \in F^Y$ be finite. Let I be the set of index pairs (i, j) such that $t_{y_i}(x_j)$ is finite, $1 \leq i \leq n$, $1 \leq j \leq m$. Then a sufficient condition that the equation $a \boxtimes t = b$ be soluble is that some solution $\{ \xi_{ij} \mid (i, j) \in I \}$ of the following optimization problem in the variables z_{ij} , for $(i, j) \in I$.

$$\begin{aligned} &\text{Minimize} && \sum_{(i,j) \in I} (b(y_i) - t_{y_i}(x_j)) z_{ij} \\ &\text{Subject to} && \left(\sum_{\substack{i=1 \\ (i,j) \in I}}^m z_{ij} \right) = 1, \quad j = 1, \dots, m \\ &\text{and} && z_{ij} \geq 0, \quad (i, j) \in I \end{aligned}$$

$$\text{...all also satisfy: } \left(\sum_{\substack{j=1 \\ (i,j) \in I}}^m \xi_{ij} \right) > 0, \quad i = 1, \dots, n.$$

We now make a definition which will be used in the next section. Let $F_{\pm\infty}$ be a belt, and let $t \in (F_{\pm\infty}^X)^Y$ be arbitrary. The right column space of t is the set of all $b \in F_{\pm\infty}^X$ for

which the equation

$$a \boxtimes t = b \quad (\text{or} \quad a \oslash t = b)$$

is soluble for a .

3.2.5. Linear Dependence

Linear dependence over a bounded l-group. We can consider the equation $a \boxtimes t = b$ (or $a \oslash t = b$) in another way. For the images t'_{x_i} , rewrite $a \boxtimes t = b$ as

$$\bigvee_{j=1}^m [t'_{x_j} \boxtimes a(x_j)] = b, \quad (3-7)$$

where $a(x_j) \in (F_{\pm\infty}^Y)^Y$ is the one-point template with target pixel value of $a(x_j)$. In this case, we say that b is a *linear combination* of $\{t'_{x_1}, t'_{x_2}, \dots, t'_{x_m}\}$, or, that $b \in F_{\pm\infty}^X$ is (*right*) *linearly dependent* on the set $\{t'_{x_1}, t'_{x_2}, \dots, t'_{x_m}\}$. We can make analogous definitions for the case of \oslash . While in linear algebra the concept of linear dependence provides a foundation for a theory of rank and dimension, the situation in the minimax algebra is more complicated. The notion of *strong linear independence* is introduced to give us a similar construct.

Theorem 3.23. *Let $F_{\pm\infty}$ be a bounded l-group other than F_3 . Let X be a coordinate set such that $|X| > 2$, and $k > 1$ be an arbitrary integer. Then we can always find k finite images on X , no one of which is linearly dependent on the others.*

If $F_{\pm\infty} = F_3$, then we can produce a dimensional anomaly.

Theorem 3.24. *Suppose $F_{\pm\infty} = F_3$, and let X be a coordinate set such that $|X| = m > 2$. Then we can always find at least $(m^2 - m)$ images on X , no one of which is linearly dependent on the others.*

Since every bounded l-group contains a copy of F_3 , the dimensional anomaly in Theorem 3.24 extends to any arbitrary bounded l-group.

Let $|X| = m$, $|Y| = n$, and $t \in (F^X)^Y$ where F is an arbitrary bounded l-group. We would like to define the rank of t in terms of linear independence, and to be equal to the number of linearly independent images t'_x of t . Suppose we were to define linear independence as the negation of linear dependence, that is, a set of k images on X (a_1, \dots, a_k) is linear independent if and only if no one of the a_i is linearly dependent on any subset of the others. Then applying Theorem 3.23 for $|X| = n$ and $k \geq n$, we could find k finite images which are linearly independent. If we defined rank as the number of linearly independent images t_y of t , then every template would have rank $k \geq n$, which is not a useful definition in this context.

Strong linear independence. As for the matrix algebra, we define the concept of strong linear independence [39].

Let $F_{\pm\infty}$ be a bounded l-group and let $a(1), \dots, a(k) \in F_{\pm\infty}^X$, $k \geq 1$. We say that the set $\{a(1), \dots, a(k)\}$ is *strongly linearly independent*, or simply SLI, if there is at least one finite image $b \in F^X$ which has a unique expression of the form

$$b = \bigvee_{p=1}^h a(j_p) \boxtimes \lambda_{j_p} \quad (\text{or} \quad b = \bigvee_{p=1}^h a(j_p) \otimes \lambda_{j_p}) \quad (3-8)$$

with $\lambda_{j_p} \in F$, $p = 1, \dots, h$, $1 \leq j_p \leq k$, $p = 1, \dots, h$, and $j_p < j_q$ if $p < q$.

If $A = \{a_1, a_2, \dots, a_k\}$ is a set of k images where each $a_i \in F_{\pm\infty}^Y$, $|Y| = n$, then we define the *template based on the set A* in the following way. For the integer k , we find a coordinate set W which has k pixel locations, that is, $|W| = k$. To this end, choose a positive integer p such that $k = p \cdot q + r$, where $r < p$ (by the division algorithm for integers). Let W denote the set $\{(i, j) : 0 \leq i \leq p-1, 0 \leq j \leq q-1\} \cup \{(-1, j) : 0 \leq j \leq r-1\}$, which is a subset of Z^2 that is almost rectangular. There is an additional row in the fourth quadrant

corresponding to the r left-over pixel locations that don't quite make a full row. Of course, there are other selections that can be made for W . Define the *template* t based on A by $t \in (F_{\pm\infty}^W)^Y$, where

$$t'_{w_i} = a_i, i = 1, \dots, k.$$

To clarify notation, we will denote the template based on the set $A = \{a_1, a_2, \dots, a_k\}$ by $t = B(A)$. Hence, if $t \in (F^X)^Y$, then for $A = \{t'_{x_1}, t'_{x_2}, \dots, t'_{x_m}\}$, we have $B(A) = t$. If $D = \{a_1, a_2, \dots, a_h\}$ is a set of h F valued images on X , we denote the *right column space* of $B(D)$ by $\langle a_1, a_2, \dots, a_h \rangle$. Thus, for $t \in (F^X)^Y$, $\langle t'_{x_1}, t'_{x_2}, \dots, t'_{x_m} \rangle$ is the right column space of t . The set $\langle a_1, a_2, \dots, a_h \rangle$ is also called the *space generated by the set* $\{a_1, a_2, \dots, a_h\}$.

Lemma 3.25. *Let $F_{\pm\infty}$ be a bounded l -group with group F . Let $c_1, \dots, c_k, b \in F_{\pm\infty}^X$, $k \geq 1$ be such that b is finite and has a unique expression of the form (3-8). Then $h = k$; $j_1 = 1, \dots, j_h = k$; $\lambda_{j_p} \in F$, $p = 1, \dots, h$; and t is doubly F -astic, where $t \in (F_{\pm\infty}^X)^Y$ is the template based on the set $C = \{c_1, \dots, c_k\}$. Here, $|Y| = k$.*

We also have

Corollary 3.26. *Let $F_{\pm\infty}$ be a bounded l -group and let $c_1, \dots, c_n \in F_{\pm\infty}^X$ for an integer $n \geq 1$. Then $\{c_1, \dots, c_n\}$ is SLI if and only if there exists a finite image $b \in F^X$ such that the equation $a \boxtimes t = b$ ($a \odot t = b$) is uniquely soluble for a , where $t \in (F_{\pm\infty}^X)^Y$ is the template based on the set $C = \{c_1, \dots, c_n\}$, $t = B(C)$, $|Y| = n$.*

We can now define linear independence. Let $F_{\pm\infty}$ be a given belt. Then *linear independence* is the negation of linear dependence: $c_1, \dots, c_n \in F_{\pm\infty}^X$ are linearly independent when no one of them is linearly dependent on the others. How is linear dependence related to strong linear independence?

Theorem 3.27. Let $F_{\pm\infty}$ be a bounded l-group, and $c_1, \dots, c_n \in F_{\pm\infty}^X$. For c_1, \dots, c_k to be linearly independent it is sufficient, but not necessary, that c_1, \dots, c_k be SLI.

We may call the above definition of SLI *right SLI*. If, in the definition of SLI, we were to multiply by the scalars λ_j 's from the left, we define the concept of *left SLI*. If formula (3-8) is replaced by

$$b = \bigwedge_{p=1}^h a(j_p) \boxtimes \lambda_{j_p} \quad (\text{or } b = \bigwedge_{p=1}^h a(j_p) \otimes \lambda_{j_p})$$

then we have the concept of *right dual SLI*. We define in an analogous way the concept of *left dual SLI*.

3.3. Rank of Templates

Template rank over a bounded l-group. Let $F_{\pm\infty}$ be a bounded l-group and $t \in (F_{\pm\infty}^X)^X$ be arbitrary. We call the template t (*right*) or *left column regular* if the set of images $\{t'_x\}_{x \in X}$ are (right) or left SLI, respectively. We say t is *right* or *left row regular* if the template t' is right or left column regular, respectively.

Now suppose that $F_{\pm\infty}$ is a bounded l-group and $t \in (F_{\pm\infty}^X)^Y$. Suppose r is the maximum number of images t'_x of t that are SLI. In this case we say that t has *column rank* equal to r . The *row rank* of t is the column rank of t' . For a template $t \in (F_{\pm\infty}^X)^Y$, we say that t has *ϕ -astic rank* equal to $r \in \mathbb{Z}^+$ if the following is true for $k = r$ but not for $k > r$:

Let W be a coordinate set, $|W| = k \leq \min(m, n)$. There exist $a \in F^X$ and $b \in F^Y$, both finite, such that the template $s \in (F_{\pm\infty}^X)^Y$ is doubly ϕ -astic and s contains a strictly doubly ϕ -astic template $u \in (F_{\pm\infty}^W)^W$, where

$$s_{y_i}(x_j) = b(y_i) + t_{y_i}(x_j) + a(x_j), \quad \forall i = 1, \dots, n \text{ and } j = 1, \dots, m$$

if $F = \mathbb{R}$, and

$$s_{y_i}(x_j) = b(y_i) * t_{y_i}(x_j) * a(x_j), \forall i = 1, \dots, n \text{ and } j = 1, \dots, m$$

if $F = R^+$.

Lemma 3.28. *Let $F_{\pm\infty}$ be a bounded l-group with group $F \in \{R, R^+\}$, and suppose that $t \in (F_{\pm\infty}^X)^Y$ has ϕ -astic rank equal to r . Then t is doubly F -astic and t' contains a set of at least r images, $t'_{x_k}, k=1, \dots, r$, which are SLI.*

Lemma 3.29. *Let $F \in \{R, R^+\}$, and suppose that $t \in (F_{\pm\infty}^X)^Y$ is doubly F -astic and consists of a set of r images which are SLI. Then t has ϕ -astic rank equal to at least r .*

Accordingly, we have

Theorem 3.30. *Let $F \in \{R, R^+\}$, and suppose that $t \in (F_{\pm\infty}^X)^Y$ is doubly F -astic. Then the following statements are all equivalent:*

- (i) t has ϕ -astic rank equal to r
- (ii) t has right column rank equal to r
- (iii) t has left row rank equal to r
- (iv) t^* has dual right column rank equal to r
- (v) t^* has dual left row rank equal to r .

If t is doubly F -astic, then we can apply Theorem 3.30 and simply use the term *rank* of t , for ranks (i) to (iii), and the term *dual rank* of t for ranks (iv) and (v). If the bounded l-group $F_{\pm\infty}$ is commutative, as in both our cases, we have the following

Corollary 3.31. *Let $F \in \{R, R^+\}$, and let $t \in (F_{\pm\infty}^X)^Y$ be doubly F -astic. Then the following statements are all equivalent:*

- (i) t has left column rank equal to r
- (ii) t has right row rank equal to r
- (iii) t^* has dual left column rank equal to r
- (iv) t^* has dual right row rank equal to r .

3.3.1. Existence of Rank and Relation to SLI

We now discuss the existence of the rank of a template and the relationship of rank to SLI.

Theorem 3.32. *Let $F \in \{R, R^+\}$, and let $t \in (F_{\pm\infty}^X)^Y$. Then there is an integer r such that t has ϕ -astic rank r , if and only if t is doubly F -astic. In this case, r satisfies $1 \leq r \leq \min(m, n)$, where $m = |X|$, $n = |Y|$.*

We now have the tools to show that the previous dimension anomalies are avoided in context of strong linear independence.

Theorem 3.33. *Let $F \in \{R, R^+\}$, X an arbitrary non-empty, finite coordinate set with $|X| = m$. Then for each integer n , $1 \leq n \leq m$, we can find n images on X , $a_j \in F_{\pm\infty}^X$, $j = 1, \dots, n$ which are SLI. This is impossible for $n > m$.*

3.3.2. Permanents and Inverses

As in linear algebra, if t is a matrix all of whose eigenvalues satisfy $|\lambda| < 1$, then the expression

$$(e - t)^{-1} = e + t + t^2 + \dots$$

is valid. We state an analogous case in the image algebra.

For a bounded l -group $F_{\pm\infty}$, a template $t \in (F_{\pm\infty}^X)^X$ is called *increasing* if

$$a \sqcap t \geq a \text{ for all } a \in F_{\pm\infty}^X, \text{ and } s \sqcap t \geq s \text{ for all } s \in (F_{\pm\infty}^X)^Y,$$

where Y is any arbitrary coordinate set.

We have

Lemma 3.34. *Let $F_{\pm\infty}$ be a bounded l -group, and let $t \in (F_{\pm\infty}^X)^X$. Then t is increasing if and only if $t_x(x) \geq \phi \forall x \in X$.*

Let $t \in (R_{\pm\infty}^X)^X$ be a template, $|X| = m$. We define the *permanent of t* to be the scalar $\text{Perm}(t) \in R_{\pm\infty}$ given by

$$\text{Perm}(t) = \bigvee_{\sigma} \left(\sum_{i=1}^m t_{y_i}(x_{\sigma(i)}) \right),$$

where the maximum is taken over all permutations σ in the symmetric group S_m of order $m!$.

For the bounded l-group $R_{\pm\infty}^+$ let $t \in ((R_{\pm\infty}^+)^X)^X$ be a template, $|X| = m$. We define the *permanent of t* to be the scalar $\text{Perm}(t) \in R_{\pm\infty}^+$ given by

$$\text{Perm}(t) = \bigvee_{\sigma} \left(\prod_{i=1}^m t_{y_i}(x_{\sigma(i)}) \right),$$

where again the maximum is taken over all permutations σ in the symmetric group S_m .

The *adjugate* template of $t \in (F_{\pm\infty}^X)^X$ is the template $\text{Adj}(t)$ defined by

$$[\text{Adj}(t)]_{y_i}(x_j) = \text{Cofactor}[t]_{x_j}(y_i)$$

where $\text{Cofactor}[t]_{x_j}(y_i)$ is the permanent of the template s defined by

$$s_{y_k}(x_h) = t_{y_k}(x_h)$$

$$h = 1, \dots, j-1, j+1, \dots, m \text{ and } k = 1, \dots, i-1, i+1, \dots, m.$$

Here, $s \in (F_{\pm\infty}^W)^W$, where $|W| = m-1$. For $m = 1$, we define $\text{Adj}(t) = \phi$.

3.3.3 Graph Theory

We now present some graph theoretic tools which will be used later.

A *digraph* or *directed graph* is a pair $D = \{V, E\}$ where V is a finite set of vertices $\{1, \dots, n\}$ and $E \subset V \times V$. The set E is called the set of *edges* of D . An edge (i, j) is *directed* from i to j , and can be represented by a vector with tail at node i and head at node j .

A *graph* is a pair $G = \{V, E\}$ where V is a finite set of vertices $\{1, \dots, n\}$ and $E \subset V \times V$ such that $(i, j) \in E$ if and only if $(j, i) \in E$.

A *u-v path* in a digraph or graph is a finite sequence of vertices $u = y_0, y_1, \dots, y_m = v$ such that $(y_j, y_{j+1}) \in E$ for all $j = 0, \dots, m-1$. A *circuit* is a path with the property that $y_0 = y_m$. A *simple path* y_0, y_1, \dots, y_m is a path with distinct vertices except possibly for y_0 and y_m . A *simple circuit* is a circuit which is a simple path.

A *weighted digraph (graph)* is a digraph (graph) to which every edge (i,j) is uniquely assigned a value in $F_{\pm\infty}$. We denote the weight of the edge (i,j) by $t(i,j)$ or t_{ij} . Note that the value t_{ij} is not necessarily equal to the value t_{ji} .

We remark that if $G = \{V, E\}$ is a graph then if there exists a $u-v$ path, there exists a $v-u$ path.

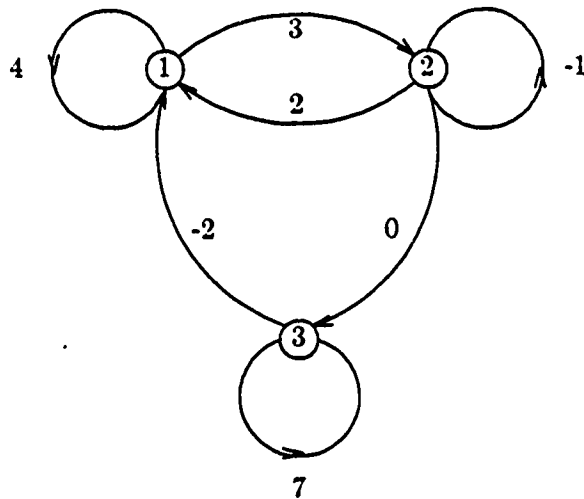
With each path (circuit) $\sigma = y_0, y_1, \dots, y_m$ of a weighted graph G , there is an associated *path (circuit) product* $p(\sigma)$, defined by

$$t_{y_0 y_1} \times t_{y_1 y_2} \times \dots \times t_{y_{m-1} y_m}.$$

For each template $t \in (F_{-\infty}^X)^X$ where $|X| = n$, we can associate a weighted graph $\Delta(t)$ in the following way. The associated graph $\Delta(t)$ is the weighted graph $G = (V, E)$, where $V = \{1, 2, \dots, n\}$, and whose weights are $t_{x_j}(x_i)$, for the pair (i, j) such that $x_i \in S_{-\infty}(t_{x_j})$. The pair (i, j) is then considered an edge. If $t_{x_j}(x_i) = -\infty$, then we can extend E to all of $V \times V$ by stating that $(i, j) \in E$ with null weight $-\infty$. An example of a template t and its associated weighted graph $\Delta(t)$ is given below in Figure 7. We have omitted listing the values of $-\infty$ on $\Delta(t)$. Here, $|X| = 3$.

$$\begin{array}{lcl}
 & & t_{x_1} = \begin{array}{|c|c|c|} \hline 4 & 2 & -2 \\ \hline \end{array} \\
 \mathbf{X} = & \begin{array}{|c|c|c|} \hline & & \\ \hline \end{array} & t_{x_2} = \begin{array}{|c|c|c|} \hline 3 & -1 & -\infty \\ \hline \end{array} \\
 & & t_{x_3} = \begin{array}{|c|c|c|} \hline -\infty & 0 & 7 \\ \hline \end{array}
 \end{array}$$

(a)



(b)

Figure 7. A Template and its Associated Graph.
 (a) A Template t ; (b) Associated Graph $\Delta(t)$.

For the belt $F_{-\infty}$, the correspondence is one-one. We note this in the next theorem.

Lemma 2.48. *Let $F_{-\infty}$ be a belt, where $+\infty \notin F$. Let $\alpha : (F_{-\infty}^X)^X \rightarrow \mathcal{G} = \{ G : G \text{ is a weighted graph with } n \text{ nodes} \}$ be defined by $\alpha(t) = \Delta(t)$. Then α is one-one and onto.*

Proof: Suppose $\alpha(t) = \alpha(s)$. Let $\{ t_{x_j}(x_i) \}$ be the weights for $\Delta(t)$ and $\{ s_{x_j}(x_i) \}$ be the weights for $\Delta(s)$. By definition, $t_{x_j}(x_i) = s_{x_j}(x_i)$ for all i, j , and hence $t = s$.

Now suppose that $G = (V, E)$ is a weighted graph with weights $\{w_{ij}\}$. Define

$t \in (F_{-\infty}^X)^X$ by $t_{x_j}(x_i) = w_{ij}$, if $(i, j) \in E$, and $t_{x_j}(x_i) = -\infty$ otherwise. Then $\alpha(t) =$

G .

Q.E.D.

Let $t \in (F_{-\infty}^X)^X$. If for each circuit σ in $\Delta(t)$ we have $p(\sigma) \leq \phi$ and there exists at least one circuit σ such that $p(\sigma) = \phi$, then we call t a *definite* template.

Lemma 3.35. *A template $t \in (F_{-\infty}^X)^X$ is definite if and only if for all simple circuits $\sigma \in \Delta(t)$, $p(\sigma) \leq \phi$ and there exists at least one such simple circuit σ such that $p(\sigma) = \phi$.*

Theorem 3.36. *Let $t \in (F_{-\infty}^X)^X$ be either row- ϕ -astic or column- ϕ -astic. Then t is definite.*

Theorem 3.37. *Let $t \in (F_{-\infty}^X)^X$. If t is definite then so is t^r , for any integer $r \geq 0$.*

Let $t \in (F_{-\infty}^X)^X$ where $|X| = n$. The *metric template* generated by t is

$$\Gamma(t) = t \vee t^2 \vee \cdots \vee t^n.$$

The *dual metric template* is

$$\Gamma^*(t) = t^* \wedge (t^2)^* \wedge \cdots \wedge (t^n)^*.$$

The name *metric* originates from the application of the minimax algebra to transportation networks. If for the bounded l-group $R_{\pm\infty}$ the value $t_{x_j}(x_i)$ represents the direct distance from node i to node j of a transportation network with $t_{x_j}(x_i) = +\infty$ if there is no direct route, then $(\Gamma(t))^*$ represents the shortest distance matrix, that is, $((\Gamma(t))^*)_{x_j}(x_i)$ is the shortest path possible from node i to node j of all possible paths. A description of a transportation problem concerning shortest paths is discussed in Cuninghame-Green's book [39].

Theorems 3.38 through 3.40 are used to prove Theorem 3.41.

Lemma 3.38. *Let $t \in (F_{-\infty}^X)^X$. Then*

$$\Gamma(t) = (1 \vee t)^{n-1} \boxtimes t.$$

Lemma 3.39. $(t \vee 1)^{n-1} = 1 \vee t \vee \dots \vee t^{n-1}$, $t \in (F_{-\infty}^X)^X$.

Theorem 3.40. *Let $t \in (F_{\pm\infty}^X)^X$ be definite. Then*

$$t^r \leq \Gamma(t), r = 1, 2, \dots$$

Theorem 3.41. *Let $t \in (F_{\pm\infty}^X)^X$ be definite. Then*

$$[\Gamma(t)]^r \leq \Gamma(t), r = 1, 2, \dots$$

$$\Gamma(t) = (1 \vee t)^r \boxtimes t, r = 1, 2, \dots, n-1.$$

Using the Adjugate of a template, we have

Theorem 3.42 [53]. *Let $F_{\pm\infty}$ be a commutative bounded l-group and $t \in (F_{\pm\infty}^X)^X$ be definite and increasing. Then $\text{Adj}(t) = \Gamma(t)$.*

Now we define the *inverse* of a template. For $t \in (F_{\pm\infty}^X)^X$, we define

$$\text{Inv}(t) = (\text{Perm}(t))^{-1} \boxtimes \text{Adj}(t) \quad (\text{or} \quad \text{Inv}(t) = (\text{Perm}(t))^{-1} \oslash \text{Adj}(t))$$

by direct analogy in elementary linear algebra.

We note that the template $\text{Inv}(t)$ is not necessarily invertible in the sense that

$$\text{Inv}(t) \boxtimes t = 1, \text{ for example.}$$

3.3.4. Invertibility

In order to define an *invertible* template, that is, a template $t \in (F_{\pm\infty}^X)^X$ that has the property that there exists a unique template s satisfying $t \boxtimes s = s \boxtimes t = 1$

$(t \oslash s = s \oslash t = 1)$, we need to introduce the concept of *equivalent templates*.

Let $F_{\pm\infty}$ be a subbelt of $R_{\pm\infty}$ or $R_{\pm\infty}^+$. A template $p \in (F_{\pm\infty}^X)^X$ is said to be *invertible* if there exists a template $q \in (F_{\pm\infty}^X)^X$ such that $p \boxtimes q = q \boxtimes p = 1$ ($p \otimes q = q \otimes p = 1$).

These templates can be described in close detail. Let us define a *strictly doubly F-astic template* over a bounded l-group $F_{\pm\infty}$ to be an element t of $(F_{\pm\infty}^X)^X$ satisfying

- (i) $t_{y_i}(x_j) < +\infty$, $i, j = 1, \dots, n$
- (ii) for each index i there exists a unique index $j_i \in \{1, 2, \dots, m\}$ such that $t_{y_i}(x_{j_i})$ is finite.

Theorem 3.43. *Let $F_{\pm\infty}$ be a bounded l-group with group F and let $p \in (F_{\pm\infty}^X)^X$ be given. Then p is invertible if and only if p is strictly doubly F-astic.*

As is usual, if p is invertible, then the template q above is written as p^{-1} .

The intersection of the set of strictly doubly ϕ -astic templates and the set of strictly doubly F -astic templates we call the *permutation templates*. It is not difficult to show

Proposition 3.44. *Let $F_{\pm\infty}$ be a bounded l-group. Then the set of invertible templates from X to X , where $|X| = m$, form a group under the multiplication \boxtimes (\otimes), containing 1 as the identity element and having the permutation templates as a subgroup isomorphic to the symmetric group S_m on m letters.*

Pre- or post-multiplication of a template t by a permutation template p will permute the images t'_x or the images t_y of t , respectively, and these permutation templates play a role exactly like their counterparts in linear algebra.

3.3.5. Equivalence of Templates

Let $F_{\pm\infty}$ be a bounded l-group, and let $t, s \in (F_{\pm\infty}^X)^Y$ be given. We say that t and s are *equivalent*, written $t \equiv s$, if there exist invertible templates $p \in (F_{\pm\infty}^Y)^Y$ and $q \in (F_{\pm\infty}^X)^X$ such that $p \boxtimes t \boxtimes q = s$ ($p \otimes t \otimes q = s$).

Now we define *elementary templates*. An *elementary template* $p \in (F_{\pm\infty}^X)^X$ over a bounded l-group with group F is one of the following:

- (i) a permutation template
- (ii) a diagonal template of the form $\text{diag}(\phi, \dots, \phi, \alpha, \phi, \dots, \phi)$, where $\alpha \in F$.

Elementary templates correspond to matrices which perform elementary operations on matrices [39]. A permutation template

- 1. permutes the images t'_x of t ; or
- 2. permutes the images t_y of t ,

depending on whether the multiplication is from the left or right, respectively. Diagonal templates of the type listed in (ii) above have the effect of multiplying some image t'_x of t by a finite constant α , or multiplying some image t_y of t by a finite constant α , depending on whether the multiplication of t is from the left or right, respectively.

Lemma 3.45. *Let $F_{\pm\infty}$ be a bounded l-group, and let X and Y be given coordinate sets, $|X|=m, |Y|=n$. Then the relation of equivalence is an equivalence relation on $(F_{\pm\infty}^X)^Y$. If $t, s \in (F_{\pm\infty}^X)^Y$, then $t \equiv s$ if and only if there is a sequence of templates u_0, u_1, \dots, u_j such that $u_0 = t$ and $u_j = s$, and u_p is obtained by an elementary operation on u_{p-1} , $p = 1, \dots, j$.*

Permutation and diagonal templates of this form will play an important role in the discussion on local template decompositions, as well as the following theorem.

Lemma 3.46. *Let $F_{\pm\infty}$ be a bounded l-group with group F and let $t \in (F_{\pm\infty}^X)^Y$ be given. If a given image of t' (or t) is F -astic then t is equivalent to a template in which that image of t' (or t) is ϕ -astic and all other images in t' (or t) are identical with the corresponding image in t' (or t). Hence if t is (row-, column-, or doubly) F -astic then t is equivalent to a template which is (respectively row-, column-, or doubly) ϕ -astic.*

Equivalence and rank. The following are results which show the relation between equivalence and rank.

Proposition 3.47. *Let $F_{\pm\infty}$ be a bounded l-group, and let $t, s \in (F_{\pm\infty}^X)^Y$. Then t has ϕ -astic rank equal to r if and only if the following statement is true for $j=r$ but not for $j > r$:*

t is equivalent to a doubly ϕ -astic template d which contains a strictly doubly ϕ -astic template $u \in (F_{\pm\infty}^W)^W$, where $|W| = j$.

Corollary 3.48. *Let $F_{\pm\infty}$ be a bounded l-group with group F and let $t, s \in (F_{\pm\infty}^X)^Y$ be equivalent. Then if either t or s has a rank, then so does the other, and the ranks are equal.*

3.4. The Eigenproblem in the Image Algebra

Using the isomorphism, we can discuss the eigenproblem which is presented in its matrix form [39] in context of the image algebra. In this section we present the eigenproblem and solution in image algebra notation.

3.4.1. The Statement in Image Algebra

Unless otherwise stated, we assume that F is a subbelt of either \mathbf{R} or \mathbf{R}^+ , and let $F_{\pm\infty}$, $F_{-\infty}$, and $F_{+\infty}$ have their usual meanings. The coordinate sets \mathbf{X} and \mathbf{Y} are assumed to be non-empty, finite arrays, with $|\mathbf{X}| = m$ and $|\mathbf{Y}| = n$.

Let $\lambda \in F_{\pm\infty}$. Let $\lambda \in (F_{\pm\infty}^X)^X$ be the one-point template defined in the usual way by

$$\lambda_y(\mathbf{x}) = \begin{cases} \lambda & \mathbf{x} = \mathbf{y} \\ -\infty & \text{otherwise} \end{cases}$$

Suppose F is a subbelt of \mathbf{R} , and $t \in (F_{\pm\infty}^X)^X$. Then the eigenproblem is to find $\mathbf{a} \in F^X$ and $\lambda \in F_{\pm\infty}$ such that

$$\mathbf{a} \boxtimes \mathbf{t} = \mathbf{a} \boxtimes \lambda.$$

Similarly for the operation \oslash , we need find $\mathbf{a} \in F^X$ and $\lambda \in F_{\pm\infty}$ such that

$$\mathbf{a} \oslash \mathbf{t} = \mathbf{a} \oslash \lambda.$$

For either belt, if such \mathbf{a} and λ exist, then \mathbf{a} is called an *eigenimage* of \mathbf{t} , and λ a corresponding *eigenvalue*. The eigenproblem is called *finitely soluble* if both \mathbf{a} and λ are finite.

As mentioned before, all results of this section are applicable for F a subbelt of with R or R^+ . Hence, to avoid stating all results for both belts separately, we will state the results for \boxtimes with the understanding that in all theorems, definitions, etc. in this section of Chapter 3, with the exception of Theorem 3.57, \boxtimes can be replaced by \oslash everywhere and the theorems and results will still hold.

Theorem 3.49. *Let $\mathbf{t} \in (F^X)^Y$. Then there exist $\mathbf{s} \in (F_{\pm\infty}^Y)^Y$ such that if \mathbf{b} is in the column space of \mathbf{t} , then \mathbf{b} is an eigenimage of \mathbf{s} with corresponding eigenvalue ϕ . Here, $\mathbf{s} = \mathbf{t}^* \boxtimes \mathbf{t} \in (F_{-\infty}^X)^X$. Hence, $\mathbf{b} \boxtimes \mathbf{t} = \mathbf{b} \boxtimes \mathbf{1} = \mathbf{b}$.*

Theorem 3.50. *Let $\mathbf{t} \in (F_{-\infty}^X)^X$. If the eigenproblem for \mathbf{t} is finitely soluble, \mathbf{t} must be row- F -astic. In particular, if \mathbf{t} is row- ϕ -astic, then the eigenproblem for \mathbf{t} is finitely soluble, in which case $\lambda = \phi$.*

Let $\mathbf{t} \in (F_{\pm\infty}^X)^X$ be definite. We know that $\Delta(\mathbf{t})$ has at least one circuit σ such that $p(\sigma) = \phi$. An *eigennode* of $\Delta(\mathbf{t})$ is any node on such a circuit. Two eigennodes are *equivalent* if they are both on any one such circuit.

Lemma 3.51. *Let $\mathbf{t} \in (F_{\pm\infty}^X)^X$ be definite. Then $\Gamma(\mathbf{t})$ is definite, and if j is an eigennode of $\Delta(\mathbf{t})$, then*

$$(\Gamma(\mathbf{t}))_{x_j}(x_j) = \phi.$$

Conversely, if $(\Gamma(\mathbf{t}))_{x_j}(x_j) = \phi$ for some $x_j \in X$, then j is an eigennode of $\Delta(\mathbf{t})$.

Lemma 3.52. Let $t \in (F_{\pm\infty}^X)^X$ be definite. If j is an eigennode of $\Delta(t)$ then

$$a^j \boxtimes t = a^j \boxtimes 1 = a^j$$

where a^j is the image $[\Gamma(t)]'_{x_j}$.

Thus, images $[\Gamma(t)]'_{x_j}$ where j is an eigennode give us eigenimages for the template t , with corresponding eigenvalue ϕ . For a given t , the set of all such images are called the *fundamental eigenimages* for t . Just as in the case for matrices, two fundamental eigenimages are called *equivalent* if nodes j and h are equivalent, and otherwise the eigenimages are non-equivalent.

Theorem 3.53. Let $t \in (F_{\pm\infty}^X)^X$ be definite. If $a^j, a^k \in F_{\pm\infty}^X$ are fundamental eigenvectors of t corresponding to equivalent eigennodes j and k , respectively, then

$$a^j = a^k \boxtimes \alpha,$$

where $\alpha \in F$, and $\alpha \in (F_{\pm\infty}^X)^X$ is the one-point template.

3.4.2. Eigenspaces

If $t \in (F_{\pm\infty}^X)^X$ is definite, let $\{a^{j_1}, \dots, a^{j_k}\}$ be a maximal set of non-equivalent fundamental eigenimages of t . The space $\langle a^{j_1}, \dots, a^{j_k} \rangle$ generated by these eigenimages is called the *eigenspace* of t .

Theorem 3.54. Let $t \in (F_{\pm\infty}^X)^X$ be given. If the eigenproblem for t is finitely soluble then every finite eigenimage has the same unique corresponding finite eigenvalue λ . The template $t \boxtimes -\lambda$ is definite, and all finite eigenimages of t lie in the eigenspace of $t \boxtimes -\lambda$. The non-equivalent fundamental eigenimages which generate this space have the property that no one of them is linearly dependent on (any subset of) the others.

The unique scalar in Theorem 3.54, when it exists, is called the *principal eigenvalue* of t .

We call a bounded l-group F *radicable* if for each $a \in F$ and integer $k \geq 1$, there exists a unique $f \in F$ such that $f^k = a$.

Some examples of radicable bounded l-groups are $R_{\pm\infty}$, $Q_{\pm\infty}$, and $R_{\pm\infty}^+$. However, $Z_{\pm\infty}$ is not radicable. Choosing $a = 12$ and $k = 5$, solving for f in the equation

$$f^5 = 12$$

is just solving for f in (using regular arithmetic)

$$5f = 12$$

which, of course, has no integral solution.

Let F be a radicable bounded l-group, and $t \in (F_{\pm\infty}^X)^X$. Let $\sigma = y_0, y_1, \dots, y_m$ be a circuit in $\Delta(t)$. We define the *length* of σ to be m . For each circuit σ in $\Delta(t)$, of length $l(\sigma)$ and having circuit product $p(\sigma)$, we define a *circuit mean* $\mu(\sigma) \in F$ by

$$[\mu(\sigma)]^{l(\sigma)} = p(\sigma).$$

We also define

$$\lambda(t) = \vee \{ \mu(\sigma) : \sigma \text{ is a simple circuit in } \Delta(t) \}.$$

For the template and associated graph $\Delta(t)$ in Figure 8, we have the following computations.

Simple Circuit σ	$p(\sigma)$	$l(\sigma)$	$\mu(\sigma)$
(1,1)	4	1	4
(2,2)	-1	1	-1
(3,3)	7	1	7
(1,2,1)	5	2	5/2
(2,3,2)	$-\infty$	2	$-\infty$
(3,1,3)	$-\infty$	2	$-\infty$
(1,2,3,1)	1	3	1/3
(3,2,1,3)	$-\infty$	3	$-\infty$

Figure 8. Computation of the Circuit Mean $\mu(\sigma)$.

In this example, $\lambda(t) = 7$.

3.4.3. Solutions to the Eigenproblem

We now present the relation between the parameter $\lambda(t)$ and the principal eigenvalue for t .

Theorem 3.55. *Let $F_{\pm\infty}$ be a radicable bounded l -group and let $t \in (R_{\pm\infty}^X)^X$ be given. If the eigenproblem for t is finitely soluble then $\lambda(t)$ is finite, and, in this case, $\lambda(t)$ is the only possible value for the eigenvalue in any finite solution to the eigenproblem for t . That is, $\lambda(t)$ is the principal eigenvalue of t .*

Theorem 3.56. *Let $F_{\pm\infty}$ be a radicable sub-bounded l -group of $R_{\pm\infty}$ and let $t \in (R_{\pm\infty}^X)^X$ be given. Then the eigenproblem for t is finitely soluble if and only if $\lambda(t)$ is finite and the template $B(A)$ is doubly F -astic, where $A =$*

$\{[\Gamma(t \boxtimes -\lambda(t))]_{x_1}', [\Gamma(t \boxtimes -\lambda(t))]_{x_2}', \dots, [\Gamma(t \boxtimes -\lambda(t))]_{x_k}'\}$ is a maximal set of non-equivalent fundamental eigenimages for the definite template $t \boxtimes -\lambda(t)$.

The Computational Task. If $|X|$ is large, and $t \in (F_{\pm\infty}^X)^X$, then to directly evaluate the circuit product for all simple circuits in t is very time consuming. We now state a theorem which makes the task more manageable for the case where the bounded l-group is $R_{\pm\infty}$.

Theorem 3.57. *Let $t \in (F_{\pm\infty}^X)^X$ be given. If the eigenproblem for t is finitely soluble, then $\lambda(t)$ is the optimal value of λ in the following linear programming problem in the $n+1$ real variables λ, x_1, \dots, x_n :*

$$\text{Minimize } \lambda \quad \text{Subject to } \lambda + x_i - x_j \geq t_{x_i}(x_j)$$

where the inequality constraint is taken over all pairs i, j for which $t_{x_i}(x_j)$ is finite.

In Theorem 3.54, we noted the linear independence of the fundamental eigenimages which generate an eigenspace. We are able now to prove a stronger result which has applications to $R_{\pm\infty}$ and $R_{\pm\infty}^+$.

Theorem 3.58 *Let $F_{\pm\infty}$ be a radicable bounded l-group other than F_3 , and let $t \in (F_{\pm\infty}^X)^X$ have a finitely soluble eigenproblem. Then the fundamental eigenimages of $-\lambda(t) \boxtimes t$ corresponding to a maximal set of non-equivalent eigennodes in $\Delta[-\lambda(t) \boxtimes t]$ are SLI.*

We now present a result relating $\lambda(t)$ and Inv .

Theorem 3.59. *Let $F_{\pm\infty}$ be a bounded l-group and $t \in (F_{\pm\infty}^X)^X$ be such that $\lambda(t) \leq \phi$. Then*

$$\text{Inv}(1 \vee t) = 1 \vee t \vee t^2 \vee \dots \vee t^K$$

for arbitrary large K . Here, 1 denotes the identity template of $(F_{\pm\infty}^X)^X$.

CHAPTER 4 GENERALIZATION OF MATHEMATICAL MORPHOLOGY

Up until the mid 1960's, the theoretical tools of quantitative microscopy as applied to image analysis were not based on any cohesive mathematical foundation. It was G. Mathéron and J. Serra at the École des Mines de Paris who first pioneered the theory of mathematical morphology as a first attempt to unify the underlying mathematical concepts being used for image analysis in microbiology, petrography, and metallography [16,53,54]. Initially its main use was to describe boolean image processing in the plane, but Sternberg [55] extended the concepts in mathematical morphology to include gray valued images via the cumbersome notion of an *umbra*. While others including Serra [56,57] also extended morphology to gray valued images in different manners, Sternberg's definitions have been used more regularly, and, in fact, are used by Serra in his book [16].

The basis on which morphological theory lies are the two classical operations of Minkowski addition and Minkowski subtraction from integral geometry [13,14]. For any two sets $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^n$, Minkowski addition and subtraction are defined as

$$A \times B = \bigcup_{b \in B} A_b \quad \text{and} \quad A / B = \bigcap_{b \in B'} A_b,$$

respectively, where $A_b = \{a + b : a \in A\}$ and $B' = \{-b : b \in B\}$. We have used the original notation as found in Hadwiger's book [14]. It can be shown that

$$A / B = (A^c \times B')^c,$$

where A^c denotes the complement of A in \mathbb{R}^n . From these definitions are constructed the two morphological operations of dilation and erosion. As used by Serra and Maragos [16,21],

the *dilation* of a set $A \subset \mathbb{R}^n$ by a structuring element $B \subset \mathbb{R}^n$ is denoted by $A \boxplus B'$ and defined by

$$A \boxplus B' = \bigcup_{b \in B'} A_b$$

while *erosion* of A by B is

$$A \boxminus B = \bigcap_{b \in B'} A_b = (A^c \boxplus B)^c.$$

We remark that the actual symbols used in Serra's and Maragos' papers for the dilation and erosion are \oplus and \ominus . To avoid confusion with the image algebra operation \oplus , we have replaced \oplus and \ominus with \boxplus and \boxminus respectively.

To avoid anomalies without practical interest, the structuring element B is assumed to include the origin $\mathbf{0} \in \mathbb{R}^n$, and both A and B are assumed to be compact. Unfortunately, the definitions for dilation and erosion defined by Serra are not the same as the Minkowski operations. In addition, while Maragos uses the same definitions as Serra for dilation and erosion, Maragos [21] uses the identical symbols \boxplus and \boxminus when defining Minkowski addition and subtraction. To add to the confusion, Sternberg defines an erosion and dilation using the same symbols \boxplus and \boxminus which are exactly the Minkowski operations [59]. The following table lists the three definitions. In all cases, $A_b = \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in A\}$, $B' = \{-\mathbf{b} : \mathbf{b} \in B\}$, and A^c denotes the complement of A in \mathbb{R}^n .

Table 2.

Minkowski	addition $A \times B = \bigcup_{b \in B} A_b$	subtraction $A / B = \bigcap_{b \in B'} A_b = (A^c \boxplus B')^c$
Serra Maragos	dilation of A by B $A \boxplus B' = \bigcup_{b \in B'} A_b$	erosion of A by B $A \boxminus B' = \bigcap_{b \in B'} A_b = (A^c \boxplus B)^c$
Sternberg	dilation of A by B $A \boxplus B = \bigcup_{b \in B} A_b$	erosion of A by B $A \boxminus B = \bigcap_{b \in B'} A_b = (A^c \boxplus B')^c$

Thus we see that while Sternberg's dilation of A by B is exactly Minkowski's addition of A and B, Serra's dilation of A by B is Minkowski's addition of A and B'. Although both definitions of erosion of A by B are equivalent to Minkowski's subtraction of A and B, Serra uses the symbol B' while Sternberg uses simply B. For the remainder of this chapter we will use Sternberg's definitions of dilation and erosion.

All morphological transformations are combinations of dilations and erosions, such as the *opening* of A by B, denoted by $A \circ B$,

$$A \circ B = (A \boxminus B) \boxplus B$$

and the *closing* of A by B, denoted by $A \bullet B$,

$$A \bullet B = (A \boxplus B) \boxminus B.$$

However, a more general image transform in mathematical morphology is the *Hit or Miss transform* [55,54]. Since an erosion and hence a dilation is a special case of the Hit or Miss

transform, this transform is often viewed as the universal morphological transformation upon which the theory of mathematical morphology is based. Let $B = (D, E)$ be a pair of structuring elements. Then the Hit or Miss transform of the set A is given by the expression

$$A \odot B = \{ a : D_a \subset A, E_a \subset A^c \}.$$

For practical applications it is assumed that $D \cap E = \emptyset$. The erosion of A by D is obtained by simply letting $E = \emptyset$, in which case we have $A \odot B = A \ominus D$.

While there have been several extensions of the boolean dilation to the gray level case, Sternberg's formulae for computing the gray value erosion and dilation are the most straightforward, although the underlying theory introduces the somewhat extraneous concept of an *umbra*. Let $f: \mathbf{R}^n \rightarrow \mathbf{R}$ be a function. Then the *umbra* of f , denoted by $\mathcal{U}(f)$, is the set $\mathcal{U}(f) \subset \mathbf{R}^{n+1}$ defined by

$$\mathcal{U}(f) = \{ p = (x, z) \in \mathbf{R}^{n+1} : z \leq f(x) \}.$$

Again, the notion of an unbounded set is exhibited in this definition, for in general the value z can approach $-\infty$. Since $\mathcal{U}(f) \subset \mathbf{R}^k$, the dilation of two functions f and g is defined through the dilation of their umbras,

$$\mathcal{U}(f \boxplus g) = \mathcal{U}(f) \boxplus \mathcal{U}(g),$$

and similarly the erosion of f by g ,

$$\mathcal{U}(f \boxminus g) = \mathcal{U}(f) \boxminus \mathcal{U}(g).$$

Any function $d: \mathbf{R}^n \rightarrow \mathbf{R}$ has the property that $d(x) = \max \{ z \in \mathbf{R} : (x, z) \in \mathcal{U}(d) \}$, and thus the set $\mathcal{U}(f \boxplus g)$ well-defines the function $f \boxplus g$. However, when actually calculating the new functions $d = f \boxplus g$ and $e = f \boxminus g$, Sternberg gives the following formulae for the two-dimensional dilation and erosion, respectively:

$$d(x,y) = \max_{i,j} [f(x-i, y-j) + g(i,j)] \quad (4-1)$$

$$e(x,y) = \min_{i,j} [f(x-i, y-j) - g(-i,-j)] \quad (4-2)$$

The function f represents the image, and g represents the structuring element. Both f and g are assumed to have finite support, with values of $-\infty$ outside. Also, in general the support of g is much smaller than the coordinate set \mathbf{X} , and $g(\mathbf{0}) \neq -\infty$. So in practice, the notion of an umbra need not be introduced at all.

Note that when applying these transforms to real data, we cannot simply substitute an image \mathbf{a} for the set A , as the expression A^c becomes meaningless to a computer. What is actually assumed is that A corresponds to the black pixels in a boolean image \mathbf{a} , that is, given $A \subset \mathbf{R}^n$, a coordinate set $\mathbf{X} \subset \mathbf{R}^n$ is chosen and a two-valued image \mathbf{a} on \mathbf{X} is found, where 1 and 0 represent the two values:

$$\mathbf{a}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \subset \mathbf{X} \\ 0 & \text{otherwise} \end{cases}$$

For the two-dimensional gray value case, Sternberg's formulas (4-1) and (4-2) are easily written in computer code, and this is, in fact, close to the image algebra definition for dilation. In short, when implementing a problem which is posed in morphological terms, the solution must be reposed in a setting which more closely represents the computing environment. On the other hand, it has been established that the image algebra comes very close to ideally modeling a large number of important image processing problems, such as mapping of transforms to sequential and parallel architectures [45] and this dissertation, and expressing sequential algorithms in a parallel manner [60].

The next part of this chapter is devoted to establishing an isomorphism between the morphological algebra and the image algebra. We will show that performing a dilation is equivalent to calculating

$$a \boxtimes t$$

for the appropriate a and t , and performing an erosion is equivalent to calculating

$$a \boxtimes t^*$$

for appropriate a and t .

Let A, B be finite subsets of \mathbb{Z}^n , where B is a structuring element. Let $X = \mathbb{Z}^n$ or choose $X \subset \mathbb{Z}^n$ to be a finite set such that $A \boxplus B \subset X$. Let F_4 denote the value set $\{-\infty, 0, 1, +\infty\}$. Define $\xi: 2^{\mathbb{Z}^n} \rightarrow F_4^X$ by $\xi(A) = a$ where

$$a(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathcal{B} = \{B \subset \mathbb{Z}^n : |B| < \infty \text{ and } 0 \in B\}$, and let I be the set of all F_4 valued invariant templates from X to X such that $y \in S_{-\infty}(t_y)$. Define $\eta: \mathcal{B} \rightarrow I$ by $\eta(B) = t$ where

$$t_y(x) = \begin{cases} 0 & \text{if } x \in B'_y \\ -\infty & \text{otherwise} \end{cases}$$

Lemma 4.1. Let ξ, η be as above. Let $A \subset \mathbb{Z}^n$, and $B \in \mathcal{B}$ a structuring element. Then

$$\xi(A \boxplus B) = \xi(A) \boxtimes \eta(B).$$

Proof: Choose X large enough such that $A \boxplus B \subset X$. Let $D = A \boxplus B$ and

$f = \xi(A) \boxtimes \eta(B)$. We must show that $y \in D$ if and only if $f(y) = 1$. To this end,

we note that

$$\begin{aligned} y \in A \boxplus B &\iff y \in A_b \text{ for some } b \in B \iff y = x + b \text{ for } x \in A, b \in B \\ &\iff x = (-b) + y, -b \in B', x \in A \iff x \in A \text{ and } x = (-b) + y \in B'_y \\ &\iff a(x) = 1 \text{ and } t_y(x) = 0 \iff \bigvee_{z \in X} a(z) + t_y(z) = f(y) = 1. \end{aligned}$$

Q.E.D.

We call a the *image corresponding to* A , and t the *template corresponding to the structuring element* B .

The next lemma shows the correspondence between the \boxminus operation and erosion.

Lemma 4.2. *Let ξ, η be as above. Let $A \subset \mathbb{Z}^n$, and $B \subset \mathbb{Z}^n$ a structuring element. Then*

$$\xi(A \boxminus B) = \xi(A) \boxminus [\eta(B)]^*.$$

Proof: Let $D = A \boxminus B$ and let $c = \xi(A) \boxminus [\eta(B)]^*$. We must show that $y \in D$ if and only if $c(y) = 1$.

$$y \in D \iff y \in A_p \forall p \in B' \iff y = x_p + p \forall p \in B',$$

where the choice of $x_p \in A$ depends on p . Let $a = \xi(A)$ and $t = \eta(B)$. Then

$$c = a \boxminus t^* \text{ and}$$

$$c(y) = \bigwedge_{x \in X} a(x) + t_y^*(x) = \bigwedge_{x \in S_{+\infty}(t_y^*)} a(x) + t_y^*(x).$$

We have

$$t_y^*(x) = [t_x(y)]^* = \begin{cases} 0 & \text{if } y \in B'_x \\ +\infty & \text{otherwise} \end{cases}$$

We claim that $S_{+\infty}(t_y^*) = B_y$. To show this, note that

$$x \in S_{+\infty}(t_y^*) \iff t_y^*(x) = 0 = t_x(y) \iff y \in B'_x \iff$$

$$y = p + x \text{ for some } p \in B' \iff x = b + y \text{ for some } b \in B \iff x \in B_y.$$

Thus,

$$y \in D \iff y = x_p + p \forall p \in B' \iff x_b = b + y \forall b \in B, \text{ for some } x_b \in A$$

$$\iff b + y = x \in A \forall b \in B \iff B_y = S_{+\infty}(t_y^*) \subset A \text{ (by definition of } \boxminus) \iff$$

$$a(x) = 1 \forall x \in B_y \subset A \text{ and } t_y^*(x) = 0 \forall x \in B_y = S_{+\infty}(t_y^*) \iff$$

$$\bigwedge_{x \in S_{+\infty}(t_y^*)} a(x) + t_y^*(x) = 1 = c(y)$$

Q.E.D.

Lemmas 4.1 and 4.2 include not only boolean but gray level dilation and erosion. However, we now show explicitly that Sternberg's formulas (4-1) and (4-2) hold in the two dimensional case. Let $f, g : \mathbb{Z}^2 \rightarrow \mathbb{R}_{-\infty}$ be two real extended real valued functions with finite support, where f represents the image and g the structuring element. Choose \mathbf{X} to be either \mathbb{Z}^2 or a finite subset containing the support of $f \boxplus g$. Then $\xi : \mathbb{R}_{-\infty}^{\mathbb{Z}^2} \rightarrow \mathbb{R}_{-\infty}^{\mathbf{X}}$ is the identity function restricted to \mathbf{X} , that is, $\mathbf{a}(\mathbf{x}) = f(\mathbf{x})$, where $\mathbf{a} = \xi(f)$. Let B denote the support of g , $B = \{ \mathbf{x} \in \mathbf{X} : g(\mathbf{x}) \neq -\infty \}$. Let I denote the set of all $\mathbb{R}_{\pm\infty}$ valued templates from \mathbf{X} to \mathbf{X} such that $\mathbf{y} \in \mathcal{S}_{-\infty}(t_{\mathbf{y}})$ for all $\mathbf{y} \in \mathbf{X}$. Define $\eta : \mathbb{R}_{-\infty}^{\mathbb{Z}^2} \rightarrow I$ by $\eta(g) = t$ where

$$t_{\mathbf{y}}(\mathbf{x}) = \begin{cases} g(\mathbf{y} - \mathbf{x}) & \text{if } \mathbf{x} \in B'_{\mathbf{y}} \\ -\infty & \text{otherwise} \end{cases}$$

Note that if $\mathbf{x} \in B'_{\mathbf{y}}$, then $\mathbf{x} = \mathbf{p} + \mathbf{y}$ for some $\mathbf{p} \in B$, which implies that $g(-\mathbf{p}) = g(\mathbf{y} - \mathbf{x})$ is well-defined. Also, $\mathcal{S}_{-\infty}(t_{\mathbf{y}}) = B'_{\mathbf{y}}$. The formal relation between the gray-scale morphological operations and \boxplus and \boxminus are shown in the next two theorems.

Theorem 4.3. *Let f, g, B , and \mathbf{X} be as above. Then*

$$\xi(f \boxplus g) = \xi(f) \boxplus \eta(g).$$

Proof:

At location $(x, y) \in \mathbf{X} \subset \mathbb{Z}^2$,

$$(a \boxplus t)(x, y) = \bigvee_{z \in \mathbf{X}} a(z) + t_{(x, y)}(z) = \bigvee_{(i, j) \in \mathcal{S}_{-\infty}(t_{(x, y)})} a(i, j) + t_{(x, y)}(i, j),$$

while at (x, y) , $f \boxplus g = d$ has value

$$d(x, y) = \max_{(i, j) \in B} [f(x - i, y - j) + g(i, j)] = \max_{(x - i, y - j) \in B} [f(i, j) + g(x - i, y - j)].$$

Given $(x - i, y - j) = (-p_1, -p_2) \in B$, we have $(i, j) = (p_1, p_2) + (x, y) \in B'_{(x, y)}$, and hence,

$$f(i, j) + g(x - i, y - j) = a(i, j) + t_{(x, y)}(i, j) \quad \forall (i, j) \in B'_{(x, y)}.$$

Therefore,

$$\begin{aligned}
 d(x,y) &= \max_{(x-i,y-j) \in B} [f(i,j) + g(x-i,y-j)] = \max_{(i,j) \in B'_{(x,y)}} [f(i,j) + g(x-i,y-j)] \\
 &= \bigvee_{(i,j) \in S_{-\infty}(t_{(x,y)})} a(i,j) + t_{(x,y)}(i,j) = (a \boxplus t)(x,y).
 \end{aligned}$$

Q.E.D.

If the template t corresponding to the structuring element g has form

$$t_y(x) = \begin{cases} g(y-x) & \text{if } x \in B'_y \\ -\infty & \text{otherwise} \end{cases}$$

then the template t^* has form

$$t_y^*(x) = \begin{cases} -g(x-y) & \text{if } y \in B'_x \\ +\infty & \text{otherwise} \end{cases}$$

Since $t_y^*(x) \in \mathbb{R}$ if and only if $x-y \in B$ if and only if $x \in B_y$, $S_{+\infty}(t_y^*) = B_y$.

Theorem 4.4. Let f, g, B , and X be as in Theorem 4.3. Then

$$\xi(f \boxplus g) = \xi(f) \boxplus [\eta(g)]^*.$$

Proof:

At location $(x,y) \in X \subset \mathbb{Z}^2$,

$$(a \boxplus t^*)(x,y) = \bigwedge_{z \in X} a(z) + t_{(x,y)}^*(z) = \bigwedge_{(i,j) \in S_{+\infty}(t_{(x,y)}^*)} a(i,j) + t_{(x,y)}^*(i,j),$$

while at (x,y) , $f \boxplus g = e$ has value

$$e(x,y) = \min_{(i,j) \in B} [f(x-i, y-j) - g(-i, -j)] = \min_{(-x+i, -y+j) \in B} [f(i,j) - g(-x+i, -y+j)].$$

Given $(-x+i, -y+j) = (b_1, b_2) \in B$, we have $(i,j) = (b_1, b_2) + (x,y) \in B$ and hence,

$$f(i,j) - g(-x+i, -y+j) = a(i,j) + t_{(x,y)}^*(i,j) \quad \forall (i,j) \in B_{(x,y)}.$$

Therefore,

$$\begin{aligned}
 e(x,y) &= \min_{(-x+i, -y+j) \in B} [f(i,j) - g(-x+i, -y+j)] = \min_{(i,j) \in B_{(x,y)}} [f(i,j) - g(-x+i, -y+j)] \\
 &= \bigwedge_{(i,j) \in S_{+\infty}(t_{(x,y)}^*)} a(i,j) + t_{(x,y)}^*(i,j) = (a \boxplus t^*)(x,y).
 \end{aligned}$$

Q.E.D.

It is easily ascertained that each of ξ and η are one-one and onto for each of the boolean and gray level cases. The functions ξ and η therefore preserve the morphological operations, and in fact theorems 4.1 through 4.4 show that mathematical morphology is embedded into the image algebra. We condense the results in the following two expressions.

$$\begin{array}{lll} \mathbf{a} \boxplus \mathbf{t} & \text{corresponds to the dilation of } f \text{ by } g, & f \boxplus g \\ \mathbf{a} \boxdot \mathbf{t}^* & \text{corresponds to the erosion of } f \text{ by } g, & f \boxdot g \end{array}$$

The operation \odot can also be used to express a boolean dilation or erosion. Given $A \subset \mathbb{R}^n$, the image $\mathbf{a} \in \{0,1\}^X$ corresponding to A is defined as before, while for a structuring element $B \subset \mathbb{R}^n$, the template \mathbf{t} corresponding to B is defined by

$$t_y(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in B'_y \\ 0 & \text{otherwise} \end{cases}$$

The $\mathbf{a} \odot \mathbf{t}$ corresponds to the dilation of A by B , while $\mathbf{a} \odot \bar{\mathbf{t}}$ corresponds to the erosion of A by B . Here, of course,

$$\bar{t}_y(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{y} \in B'_x \\ +\infty & \text{otherwise} \end{cases}$$

Thus, the value set $\{-\infty, 0, 1, +\infty\}$ along with the operation \boxplus can be used to express a boolean dilation, where both input image \mathbf{a} and output $\mathbf{a} \boxplus \mathbf{t}$ are $\{0,1\}$ valued images. Similarly, the three element blog $F_3 = \{0,1,+\infty\}$ along with the operation \odot can be used to express a boolean dilation, where both \mathbf{a} and $\mathbf{a} \odot \mathbf{t}$ are $\{0,1\}$ valued images.

In the boolean case it is a simple exercise to show that the Hit or Miss transform can be expressed as

$$A \odot B = (A \boxdot D) \cap (A^c \boxdot E). \quad (4-3)$$

If we let t be the boolean template corresponding to D and s the boolean template corresponding to E , we obtain the equivalent image algebra expression

$$(a \boxtimes t^*) * (\chi_0(a) \boxtimes s^*), \quad (4-4)$$

or, using the bounded l-group $F_3 = \{0, 1, +\infty\}$,

$$(a \oslash \bar{t}) * (\chi_0(a) \oslash \bar{s}). \quad (4-5)$$

However, there is an even simpler image algebra formulation of the Hit or Miss transform which does not employ the notions of minimum or erosion. We state this in the next lemma, using equation (4-5) as the representation of the Hit or Miss transform. Before doing so, however, we introduce the concept of a *census template*. Let $S \subset X \subset \mathbb{R}^n$, where X is a finite array. Suppose every vector $s = (s_1, \dots, s_n) \in S$ is assigned a value of 0 or 1, and we wish to determine if the pattern of 0's and 1's within S matches some predetermined pattern we want to identify. One can uniquely identify every possible combination of 0 and 1 values within S by attaching a unique number to each combination. This is done by using an invariant template whose support is the set S and each weight $t_y(x)$, $x \in S$, is a distinct power of a prime number. When applied via the operation \oplus to a boolean image, each distinct pattern of 0's and 1's in S will result in a unique number. While this theory is general enough to identify patterns in n -dimensional space, it is primarily used for 2-dimensional image processing. The following is a 2-dimensional census template.

16	8	4
32	1	2
64	128	256

Figure 9. An Example of a Census Template.

This template was used as part of an algorithm to determine the chain code of a boolean image [60].

Lemma 4.5. *Let t be the boolean template corresponding to D and s the boolean template corresponding to E in the Hit or Miss transform (4-5). Suppose $S_{+\infty}(\bar{t}_y) = \{x_1, \dots, x_k\}$ and $S_{+\infty}(\bar{s}_y) = \{x_{k+1}, \dots, x_n\}$, where the x_i are distinct, $i = 1, \dots, n$. Then*

$$\chi_m(a \oplus r), \text{ where } m = \sum_{i=1}^k 2^{i-1},$$

is equivalent to computing the Hit or Miss transform, and the n -point census template

$r \in (R^X)^X$ is defined by

$$r_y(x_i) = \begin{cases} 2^{i-1} & i=1, \dots, n \\ 0 & x \neq x_i \text{ for any } i=1, \dots, n \end{cases}$$

Proof: Note that $a \in \{0,1\}^X$ and $\chi_m(a \oplus r) \in \{0,1\}^X$ also. Let $b = (a \oslash \bar{t}) * (\chi_0(a) \oslash \bar{s})$, and let $c = \chi_0(a)$. Then at $y \in X$, expression (4-5) has value

$$\begin{aligned} b(y) &= \left(\bigwedge_{x \in S_{+\infty}(\bar{t}_y)} a(x) * \bar{t}_y(x) \right) * \left(\bigwedge_{x \in S_{+\infty}(\bar{s}_y)} c(x) * \bar{s}_y(x) \right) \\ &= \left[\bigwedge_{i=1}^k a(x_i) * \bar{t}_y(x_i) \right] * \left[\bigwedge_{i=k+1}^n c(x_i) * \bar{s}_y(x_i) \right]. \end{aligned}$$

The pixel $b(y)$ will have value 1 if and only if each of the expressions

$$\left[\bigwedge_{i=1}^k a(x_i) * \bar{t}_y(x_i) \right] \quad (4-6)$$

$$\left[\bigwedge_{i=k+1}^n c(x_i) * \bar{s}_y(x_i) \right] \quad (4-7)$$

has value 1. Note that the only other possible value that (4-6) or (4-7) can assume is

0. Now,

$$\left[\bigwedge_{i=1}^k a(x_i) * \bar{t}_y(x_i) \right] = 1 \iff a(x) * \bar{t}_y(x) = 1 \quad \forall i = 1, \dots, k,$$

$$\iff a(x_i) = 1 \text{ and } \bar{t}_y(x_i) = 1 \quad \forall i = 1, \dots, k.$$

Also,

$$\left[\bigwedge_{i=k+1}^n c(x_i) * \bar{s}_y(x_i) \right] = 1 \iff c(x) * \bar{s}_y(x) = 1 \quad \forall i = k+1, \dots, n,$$

$$\iff c(x_i) = 1 \text{ and } \bar{s}_y(x_i) = 1 \quad \forall i = k+1, \dots, n.$$

Thus, expression (4-5) will have value 1 if and only if $a(x_i) = 1$ for all $i = 1, \dots, k$ and $c(x_i) = 1$ for all $i = k+1, \dots, n$. But $c(x_i) = 1$ if and only if $a(x_i) = 0$, and this is true for all $i = k+1, \dots, n$. Therefore (4-5) will have value 1 if and only if $a(x_i) = 1$ for all $i = 1, \dots, k$ and $a(x_i) = 0$ for all $i = k+1, \dots, n$. The image $\chi_m(a \oplus r)$ will assume a non-zero value only when $a \oplus r$ has value m . This happens if and only if $a(x_i) = 1$ for all $i = 1, \dots, k$ and $a(x_i) = 0$ for all $i = k+1, \dots, n$, as

$$\begin{aligned} (a \oplus r)(y) &= \sum_{x \in S(r_y)} a(x) \cdot r_y(x) = \sum_{i=1}^n a(x_i) \cdot r_y(x_i) = \sum_{i=1}^n a(x_i) \cdot 2^{i-1} \\ &= \begin{cases} m = \sum_{i=1}^k 2^{i-1} & \text{if and only if } a(x_i) = 1 \text{ for } i = 1, \dots, k \text{ and} \\ & a(x_i) = 0 \text{ for } i = k+1, \dots, n \\ \text{an integer } \neq m & \text{otherwise} \end{cases} \end{aligned}$$

Here we use the fact that $S(r_y) = S_{+\infty}(\bar{t}_y) \cup S_{+\infty}(\bar{s}_y)$, and also that

$S_{+\infty}(\bar{t}_y) \cap S_{+\infty}(\bar{s}_y) = \emptyset \forall y \in X$ (as $D \cap E = \emptyset$). Therefore we see that (4-5) will have value 1 at location y if and only if $\chi_m(a \oplus r)$ has value 1 at location y .

Q.E.D.

We have shown that the subalgebra $\mathcal{A} = (R^X, I, \vee, \boxminus, \wedge, \boxplus)$ of the full image algebra is isomorphic to the morphological algebra as described by Serra and Sternberg. Since invariant templates with the target pixel included in their support (the set I) are a special type of templates, it is clear that $(R^X, (R_{\pm\infty}^X)^X, \vee, \boxminus, \wedge, \boxplus)$ is a much larger algebra than the morphological algebra. Templates generalize the concept of a structuring element. Templates can vary in size, shape, and weights from point to point and they are able to express a more general mapping, taking an image with possibly values of $+\infty$ in m -dimensional space to an image in n -dimensional space if we replace $(R_{-\infty}^X)^X$ by $(R_{\pm\infty}^X)^Y$, where $X \subset R^m$ and $Y \subset R^n$. Thus, an expression of form $a \boxminus t$ can represent a far more complex process than a simple dilation. For example, let a denote the input image shown in Figure 10(a) and define t by

$$t_{(x,y)}(i,j) = \begin{cases} 0 & \text{if } (x,y) = (i,j) \text{ or} \\ & \text{if } (x,y) \in S_1 \cup S_2 \text{ and } (0,0) = (i,j) \\ -\infty & \text{otherwise} \end{cases}$$

where $S_1 = \{ (x,y) : 0.9 < \frac{x^2}{p^2} + \frac{y^2}{q^2} < 1.1 \}$ and $S_2 = \{ (x,y) : 0.9 < \frac{x^2}{d^2} - \frac{y^2}{q^2} < 1.1 \}$, $c = 30$, $q=15$, $p^2 = q^2 + c^2$, $d^2 = c^2 - q^2$. In this case $a \boxminus t$ is obviously not a simple dilation. It is not at all clear if this transformation can be expressed in terms of dilations and erosions, starting with the input image a , and if this is indeed possible, if the resulting expression would be transparent enough to justify the effort. The input and output images are shown in the next two figures.

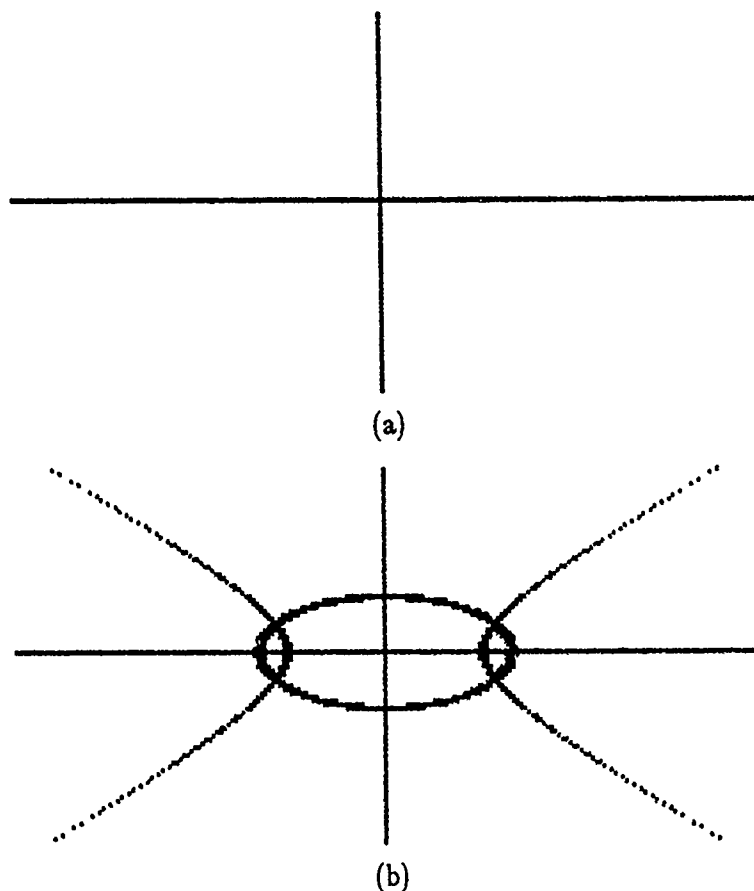


Figure 10. Example of a Non-morphological Transformation.
 (a) Input Image a ; (b) Image $a \boxtimes t$.

As a last remark, we state that a structuring element corresponds to a square matrix that is block toeplitz with toeplitz blocks that has finite elements on the diagonal. A decomposition technique for a class of invariant templates called rectangular templates is discussed in Chapter 5, and presents the method in matrix as well as image algebra notation.

CHAPTER 5 TRANSFORM DECOMPOSITION

5.1. New Matrix Decomposition Results

In this section, we state matrix results which do not appear in the book *Minimax Algebra* and which are new results. We will have particular use for most of the material presented here, as it will be used to give necessary and sufficient conditions for local decomposition of lattice transforms. The isomorphisms are used to map matrix algebra techniques to the image algebra.

As mentioned in the introduction, the use of parallel processors in computing image processing transforms is on the increase. Most transforms are not able to be applied directly to a parallel architecture. Instead, a particular transform must usually be *mapped* to a specific architecture, that is, the limitations of the machinery upon which the transform is to be implemented must be represented in the mathematical expression of the transform. For example, in the case of the neighborhood array processors, this involves *decomposing* the transform into a sequence of factors where each factor is directly implementable on the architecture. To give the general idea of this approach, we represent the set of processors by a rectangular array X , $X = \{(i,j) : 0 \leq i \leq m, 0 \leq j \leq n\}$, and if the neighborhood has a von Neumann or Moore configuration, then communication links between each processor and the neighboring processors to which it is connected (the *local network*) are represented by one of the diagrams in Figure 1.

Here, the processor x can distribute and receive information from its four (or eight) nearest neighbors. A transform t is represented by a template $t \in (F^X)^X$. If we

can write t as a product of templates,

$$t = t^1 \boxtimes t^2 \boxtimes \cdots \boxtimes t^k$$

where each t^i has its support $S_{-\infty}(t_y^i)$ as a subset of the local network configuration (in this example, the von Neumann or Moore configuration), for all $y \in X$, then we say that t has a local decomposition with respect to the network. Thus, the important consideration in determining decompositions is the underlying network of communication between processors. The network can be modeled by a graph or digraph, with nodes as the processors and edges or directed edges as the communication links between processors. The results on decomposition that follow will enable us to determine necessary and sufficient conditions on the graph of the network to guarantee the existence of a local transform decomposition under the operation of \boxtimes or \boxplus .

In this chapter, we assume that $F_{\pm\infty}$ is a sub-bounded l-group of $R_{\pm\infty}$ or $R_{\pm\infty}^+$ where F is the group of the bounded l-group $F_{\pm\infty}$, and $-\infty$ and $+\infty$ have their usual meaning in context of the respective bounded l-groups. Some sub-bounded l-groups are $R_{\pm\infty}$, $Q_{\pm\infty}$, $Z_{\pm\infty}$, and $Q_{\pm\infty}^+$. Further, we assume that any matrix described in these sections, unless noted otherwise, will assume no values of $+\infty$.

5.1.1. Preliminaries

We first give the results in matrix notation in section 5.1, and then use the isomorphism to map the results to image algebra notation in section 5.2.

Properties of the transpose and conjugate. Recall the notation t' which denotes the transpose of the matrix t , and that $(t')' = t$.

Theorem 5.1. *Let $s \in M_{mp}$, $t \in M_{pn}$ be given. Then $(s \times t)' = t' \times s'$, and $(s \times' t)' = t' \times' s'$.*

Proof: We give the proof for $(s \times t)' = t' \times s'$. The case where \times' replaces \times is done in a similar way. First, note that $(s \times t)' \in M_{nm}$, and $t' \times s' \in M_{nm}$ also. Let $r = s \times t$, and $u = t' \times s'$. To prove the lemma, it is equivalent to show that $r = u'$:

$$r_{ij} = u_{ji}.$$

We have

$$r_{ij} = \bigvee_{k=1}^n s_{ik} \times t_{kj}.$$

The elements of the j -th row of t' are $t_{1j}, t_{2j}, \dots, t_{nj}$ and the elements of the i -th column of s' are $s_{i1}, s_{i2}, \dots, s_{in}$. Thus, the element in the j -th row and i -th column of $u = t' \times s'$ is

$$u_{ji} = \bigvee_{h=1}^n t_{hj} \times s_{ih}.$$

But

$$u_{ji} = \bigvee_{h=1}^n t_{hj} \times s_{ih} = \bigvee_{h=1}^n s_{ih} \times t_{hj} = r_{ij}.$$

Q.E.D.

Theorem 5.2. $a \times' b = (b^* \times a^*)^*$, where a, b take values in $R_{\pm\infty}$ or $R_{\pm\infty}^+$.

Proof: For matrices of the appropriate sizes, we know

$$(f \times g)^* = g^* \times' f^*,$$

as M_{mn} is conjugate to M_{mn}^* [39]. Thus, setting $g = a^*$ and $f = b^*$, we have

$$(b^* \times a^*)^* = a \times' b.$$

Q.E.D.

Let $V = \{1, 2, \dots, n\}$, and let $W: V \rightarrow V$ be any function with the property $i \in W(i)$ for all i . Under the induced action of the isomorphism Ψ^{-1} as discussed in Chapter 2, W represents the neighbors of processor x_i with whom processor x_i can communicate. That is,

$\Psi^{-1}(W(i))$ represents those processors in a network whose memories processor x_i can directly access, which includes itself (as $i \in W(i)$). We shall call W a *configuration function* on V .

An example of a configuration function is the von Neumann configuration function. See Figure 1(a). As before, we assume our coordinate set \mathbf{X} is an $r \times s$ array, $rs = n$. Fix $i \in V$. Then $i = k*s + p$, for some non-negative integers k, p where $0 \leq p < s$ (by the Euclidean algorithm for the integers). Let $N(i) = \{i-1, i, i+1, i-s, i+s\}$, for $i=1, \dots, n$. Then the von Neumann configuration function W is defined by $W(i) = \{j \in N(i) : j \in V, \text{ and } j \text{ satisfies one of the 4 conditions: (1) } j = i; (2) j = h*s + p, \text{ where } h = k-1 \text{ or } h = k+1; (3) \text{ if } p \neq 0, \text{ then } j = k*s + 2; \text{ or (4) if } p \neq 1, \text{ then } j = k*s - 1\}$. This formulation takes care of the truncated von Neumann neighborhood on the boundary pixels of \mathbf{X} , as well as for pixels on the interior of \mathbf{X} .

We now state some preliminary definitions and concepts, making use of the graph theory developed in section 3.3.3.

Define $S_{-\infty}(t_i) = \{j \in V : t_{ij} \neq -\infty\}$. The set $S_{-\infty}(t_i)$ is called the *support* of t_i , in accordance with the image algebra definition of the support of a template at location x_i . In fact, $\Psi^{-1}(S_{-\infty}(t_i)) = S_{-\infty}(t'_{x_i}) \forall i$. Let W be a configuration function, and let $t \in M_{nn}$.

We say t is *local with respect to* W if $S_{-\infty}(t_i) \subset W(i)$ for all $i \in V$. If W is understood we say that t is local. A *decomposition* of t is a set of matrices $\{t(i)\}_{i=1}^j$ such that $t = \bigotimes_{i=1}^j t(i)$.

The set $\{t(i)\}_{i=1}^j$ is a *local decomposition* of $t \in M_{nn}$ with respect to W if $t(i)$ is local with respect to W for all $i=1, \dots, j$. As before, if W is understood, we say simply that $\{t(i)\}_{i=1}^j$ a local decomposition of t . A *weak decomposition* is one in which V occurs as a template operation. A *weak local decomposition* is a local decomposition that is weak.

Let $D = \{V, E\}$ be a digraph with $u, v \in V$. We say that v is *reachable from* u if there exists a path from u to v in D . If u is reachable from v and v is reachable from u , then we say that the pair (u, v) or (v, u) is *mutually reachable* in D . Note that in a graph G , reachable and mutually reachable are equivalent. A digraph or graph is *strongly connected* if for all pairs $(u, v) \in V \times V$, u is reachable from v .

We now present the correspondence between template configurations and digraphs.

For every configuration function W , we can associate a graph and digraph in the following way. For $i \in V$, let $E_i = \{(j, i) : j \in W(i)\}$, and let $F_i = \{(i, j) : (j, i) \in E_i\}$. The *digraph of* W , denoted by $D(W)$, is the digraph $\{V, E\}$ where $E = \bigcup_{i=1}^n E_i$. The *graph of* W is denoted by $G(W)$, and is defined to be $G(W) = \{V, E\}$, where $E = \bigcup_{i=1}^n E_i \cup \bigcup_{i=1}^n F_i$.

We remark that if W is the von Neumann or the Moore configuration function, then $G(W)$ is strongly connected. These are common types of neighborhood connection schemes used on parallel architectures.

Now, we establish the correspondence between our weighted graph $G(W)$ and a matrix. Recall in Chapter 3, we described a one-one correspondence between a graph and a template $t \in (F_{-\infty}^X)^X$. We use the isomorphism Ψ^{-1} to map a template to a matrix, and define the weighted graph in terms of the function α and Ψ^{-1} . Specifically, for a matrix $t = (t_{ij})$, we define the corresponding graph $\Delta(t)$ as

$$\Delta(t) \equiv \alpha(\Psi^{-1}(t)),$$

and hence $\Delta(t)$ has edge weight t_{ji} for the edge (i, j) . Let W be a configuration function, and $G(W)$ its graph. Then any matrix t associated with $G(W)$ must satisfy $t_{ji} = -\infty$ if (i, j) is not an edge in $G(W)$. We will use this fact in determining local decompositions for an arbitrary matrix t . The reason for the exchange of indices in this correspondence now becomes

clear. If $(j,i) \in E$ then $j \in W(i)$, and local matrices must satisfy the condition that $S_{-\infty}(t_i) \subset W(i)$. Thus, for a given $i \in V$, $\{j \in V : j \in W(i)\}$ represents the possible indices j where t_{ij} need not have value $-\infty$, or, equivalently, the possible processors x_j who can communicate with processor x_i .

5.1.2. Local Decompositions

A *matrix decomposition* of t is a set of matrices $t(1), \dots, t(j)$ such that $t = t(1) \times t(2) \times \dots \times t(j)$. The $t(i)$ are called the *factors* of the decomposition. We write $t = \prod_{i=1}^j t(i)$ is a decomposition of t .

Lemma 5.3. *Suppose that $s \times t = r$ is a decomposition of r . Then this decomposition is not unique, and we have $\hat{s} \times \hat{t}$ is also a decomposition of r , with*

$$\hat{s} = \lambda \times s, \quad \hat{t} = \lambda^{-1} \times t$$

and $\lambda \in F$ is arbitrary.

Proof: Suppose that $s \times t$ is a decomposition of r , and let $\lambda \in F$ be arbitrary. Then

$\lambda \times s = s \times \lambda$, as (in our case) F is commutative. This implies that

$$\begin{aligned} (\lambda \times s) \times (\lambda^{-1} \times t) &= (s \times \lambda) \times (\lambda^{-1} \times t) \\ &= s \times (\lambda \times \lambda^{-1}) \times t = s \times e \times t = s \times t = r \end{aligned}$$

Thus, setting $\hat{s} = \lambda \times s$, and $\hat{t} = \lambda^{-1} \times t$, we see that $\hat{s} \times \hat{t} = r$ also.

Q.E.D.

Lemma 5.4. *Let $t \in M_{nn}$ be such that $t_{ii} \in F \forall i = 1, \dots, n$. Then t is equivalent to a matrix s which has the property that $s_{ii} = \phi \forall i = 1, \dots, n$.*

Proof: Let $d = \text{diag}(t_{11}^{-1}, t_{22}^{-1}, \dots, t_{nn}^{-1})$. Note that d is invertible and $d^{-1} =$

$\text{diag}(t_{11}, t_{22}, \dots, t_{nn})$. Defining $s = d \times t$, we find that in computing s_{ij} ,

$$\begin{aligned}
 s_{ii} &= \bigvee_{k=1}^n d_{ik} \times t_{ki} \\
 &= d_{ii} \times t_{ii} = t_{ii}^{-1} \times t_{ii} = \phi.
 \end{aligned}$$

Hence we have

$$s = d \times t = d \times t \times e$$

which implies, since both d and e are invertible, that

$$t = d^{-1} \times s = d^{-1} \times s \times e$$

for the $n \times n$ identity matrix e , and, thus, t is similar to s which has the required form.

Q.E.D.

A matrix $t \in M_{nn}$ satisfying $t_{ii} = \phi \forall i = 1, \dots, n$ is called a ϕ -diagonal matrix, or ϕ -diagonal, for convenience of notation.

Since every matrix t such that $t_{ii} \in F$ is equivalent to one which has ϕ 's on the diagonal, we may use this to our advantage and prove our theorems for this special type of matrix if it is easier to do, and use the property of equivalence to show that the theorems hold in the more general case.

A square matrix $t \in M_{nn}$ is said to be *lower triangular* if it satisfies

$$t_{ij} = -\infty \text{ if } i < j$$

and *upper triangular* if it satisfies

$$t_{ij} = -\infty \text{ if } i > j.$$

Lemma 5.5. *Let $t \in M_{nn}$ be ϕ -diagonal. Then t has a weak decomposition into lower and upper triangular matrices. In particular, t can be written as*

$$t = l \vee u,$$

where l (u) is lower (upper) diagonal, and $l_{ii} = u_{ii} = \phi$.

Proof: Define $l, u \in M_{nn}$ by

$$l_{ij} = \begin{cases} t_{ij} & \text{if } i \leq j \\ -\infty & \text{otherwise} \end{cases}$$

$$u_{ij} = \begin{cases} t_{ij} & \text{if } i \geq j \\ -\infty & \text{otherwise} \end{cases}$$

It is easy to see that $l \vee u = t$. This is because

$$[l \vee u]_{ij} = l_{ij} \vee u_{ij} = \begin{cases} t_{ij} \vee -\infty & \text{if } i < j \\ -\infty \vee t_{ij} & \text{if } i > j \\ t_{ii} \vee t_{ii} = \phi & \text{if } i = j \end{cases}$$

which shows that $l \vee u = t$.

Q.E.D.

Corollary 5.6. Let $t \in M_{nn}$ be lower or upper triangular with the property that $t_{ii} \in F \forall i$.

Then t is equivalent to a matrix which is ϕ -diagonal.

An *off* matrix $b \in M_{nn}$ is a matrix which satisfies:

$b_{ii} \in F \forall i = 1, \dots, n$, and $b_{ij} = -\infty$ if $i \neq j$, except for a unique index pair (i', j') such that $b_{i', j'} \in F$.

If $t \in M_{mn}$, then we use the notation t_i to denote the i -th row of t , and the notation t^j to denote the j -th column of t . If $a \in E^n$, then a_i denotes the i -th entry of the vector a . If $t \in M_{mn}$, then t_{ij} denotes the entry at location (i, j) . Thus, for $t \in M_{mn}$, $(t)_j^i = t_{ij}$, and it is trivial to show

Proposition 5.7. $(s \times t)_j^i = s_i \times t^j, \neg$.

Let $l \in M_{nn}$ be lower triangular with all diagonal entries equal to ϕ . For $k=1, \dots, n$, define

$${}^k c = e \vee [l^k \times (e^k)'].$$

Thus, ${}^k c$ has form

$${}^k c = \begin{bmatrix} \phi & & & & \\ & \phi & & & \\ & & \phi & & \\ & & & l_{k+1,k} & \phi \\ & & & & \vdots \\ -\infty & & & l_{nk} & -\infty & \phi \end{bmatrix}.$$

Lemma 5.8. *If $l \in M_{nn}$ is lower triangular and ϕ -diagonal, then*

$$l = {}^1 c \times {}^2 c \times \cdots \times {}^{n-1} c.$$

Proof: By induction we show that $s = {}^1 c \times {}^2 c \times \cdots \times {}^k c$ has form

$$s = \begin{bmatrix} \phi & & & & & \\ l_{21} & \phi & & & & \\ & l_{32} & & & & \\ & & \phi & & & \\ & & & l_{k+1,k} & \phi & \\ & & & & -\infty & \phi \\ & & & & & \vdots \\ l_{n1} & l_{n2} & & l_{nk} & & -\infty & \phi \end{bmatrix} \quad (5-1)$$

for $1 \leq k \leq n-1$. It is easily shown that

$${}^1 c \times {}^2 c = \begin{bmatrix} \phi & & & & & \\ l_{21} & \phi & & & & \\ & l_{32} & & & & \\ & & \phi & & & \\ & & & -\infty & \phi & \\ & & & & -\infty & \phi \\ & & & & & \vdots \\ l_{n1} & l_{n2} & -\infty & -\infty & & -\infty & \phi \end{bmatrix}.$$

We assume that $s = {}^1 c \times {}^2 c \times \cdots \times {}^k c$ has form as in (5-1) for $1 \leq k \leq n-2$.

Assume the induction step, and consider $s \times {}^{k+1}c$.

Case 1. $j \neq k+1$. $(s \times {}^{k+1}c)^j = s \times ({}^{k+1}c)^j = s \times e^j$, as $({}^{k+1}c)^j = e^j$ for $j \neq k+1$.

$$\text{Continuing, } s \times e^j = s^j = \begin{cases} l^j & 1 \leq j \leq k \\ e_j & k+1 < j \leq n \end{cases}$$

Case 2. $j = k+1$. $(s \times {}^{k+1}c)_{i,k+1} = (s \times {}^{k+1}c)_i^{k+1} = s_i \times ({}^{k+1}c)^{k+1} =$

$$\bigvee_{h=1}^n s_{ih} \times {}^{k+1}c_{h,k+1}. \text{ If } i < k+1, \text{ then}$$

$$\begin{aligned} \bigvee_{h=1}^n s_{ih} \times {}^{k+1}c_{h,k+1} &= \left[\bigvee_{h=1}^i s_{ih} \times {}^{k+1}c_{h,k+1} \right] \vee \left[\bigvee_{h=i+1}^{k+1} s_{ih} \times {}^{k+1}c_{h,k+1} \right] \vee \\ &\left[\bigvee_{h=k+2}^n s_{ih} \times {}^{k+1}c_{h,k+1} \right] \\ &= \left[\bigvee_{h=1}^i l_{ih} \times -\infty \right] \vee \left[\bigvee_{h=i+1}^{k+1} -\infty \times {}^{k+1}c_{h,k+1} \right] \vee \left[\bigvee_{h=k+2}^n -\infty \times {}^{k+1}c_{h,k+1} \right] = -\infty. \end{aligned}$$

$$\text{If } i = k+1, \text{ then } \bigvee_{h=1}^n s_{ih} \times {}^{k+1}c_{h,k+1} =$$

$$\begin{aligned} &\left[\bigvee_{h=1}^k s_{k+1,h} \times {}^{k+1}c_{h,k+1} \right] \vee \left[s_{k+1,k+1} \times {}^{k+1}c_{k+1,k+1} \right] \vee \left[\bigvee_{h=k+2}^n s_{k+1,h} \times {}^{k+1}c_{h,k+1} \right] \\ &= \left[\bigvee_{h=1}^k l_{k+1,h} \times -\infty \right] \vee \left[\phi \times \phi \right] \vee \left[\bigvee_{h=k+2}^n -\infty \times {}^{k+1}c_{h,k+1} \right] = \phi. \end{aligned}$$

$$\text{If } i > k+1, \text{ then } \bigvee_{h=1}^n s_{ih} \times {}^{k+1}c_{h,k+1} = \left[\bigvee_{h=1}^k s_{ih} \times {}^{k+1}c_{h,k+1} \right] \vee \left[\bigvee_{h=k+1}^{i-1} s_{ih} \times {}^{k+1}c_{h,k+1} \right]$$

$$\begin{aligned} &\vee \left[s_{ii} \times {}^{k+1}c_{i,k+1} \right] \vee \left[\bigvee_{h=i+1}^n s_{ih} \times {}^{k+1}c_{h,k+1} \right] \\ &= \left[\bigvee_{h=1}^k l_{ih} \times -\infty \right] \vee \left[\bigvee_{h=k+1}^{i-1} -\infty \times {}^{k+1}c_{h,k+1} \right] \vee \left[\phi \times l_{i,k+1} \right] \vee \\ &\left[\bigvee_{h=i+1}^n -\infty \times {}^{k+1}c_{h,k+1} \right] = l_{i,k+1}. \end{aligned}$$

$$\begin{bmatrix} \phi & & & & & \\ -\infty & & & & & \\ & \phi & & & -\infty & \\ & l_{k+1,k} & \phi & & & \\ & l_{k+2,k} & -\infty & \phi & & \\ & -\infty & & & \phi & \\ -\infty & & & & & \\ & -\infty & & -\infty & -\infty & \phi \end{bmatrix}.$$

Let $\mathbf{s} = {}^{i,k}\mathbf{c} \times {}^{i-1,k}\mathbf{c} \times \dots \times {}^{k+1,k}\mathbf{c}$, for $i \leq n-1$. Here, we assume

$$\mathbf{s}^j = \begin{cases} \mathbf{e}^j & \text{if } j \neq k \\ \begin{bmatrix} -\infty \\ \cdot \\ \phi \\ l_{k+1,k} \\ l_{k+2,k} \\ \cdot \\ l_{i,k} \\ -\infty \\ \cdot \\ -\infty \end{bmatrix} & \text{if } j = k \end{cases}$$

Then the j -th column of ${}^{i+1,k}\mathbf{c} \times \mathbf{s}$ is $({}^{i+1,k}\mathbf{c} \times \mathbf{s})^j$.

Case 1. $j \neq k$. ${}^{i+1,k}\mathbf{c} \times \mathbf{s}^j = {}^{i+1,k}\mathbf{c} \times \mathbf{e}^j = {}^{i+1,k}\mathbf{c}^j = \mathbf{e}^j$.

Case 2. $j = k$.

$${}^{i+1,k}\mathbf{c}_m \times \mathbf{s}^k = \bigvee_{h=1}^n {}^{i+1,k}\mathbf{c}_{mh} \times \mathbf{s}_{hk}. \quad (5-2)$$

If $m < k$, then (5-2) is ${}^{i+1,k}\mathbf{c}_m \times \mathbf{s}^k = \mathbf{e}_m \times \mathbf{s}^k = -\infty$.

For $k \leq m \leq i$, equation (5-2) gives us ${}^{i+1,k}\mathbf{c}_m \times \mathbf{s}^k = \begin{cases} \phi & \text{if } m = k \\ l_{mk} & \text{if } m = k+1, \dots, i \end{cases}$.

If $m = i+1$ then ${}^{i+1,k}\mathbf{c}_m \times \mathbf{s}^k = {}^{i+1,k}\mathbf{c}_{i+1} \times \mathbf{s}^k = l_{i+1,k}$.

If $m > i+1$ then ${}^{i+1,k}c_m \times s^k = -\infty$.

Thus, the k -th column of ${}^{i+1,k}c \times s$ has form

$$\begin{bmatrix} -\infty \\ \cdot \\ \phi \\ l_{k+1,k} \\ \cdot \\ \cdot \\ l_{i+1,k} \\ -\infty \\ \cdot \\ -\infty \end{bmatrix}.$$

Q.E.D.

We now state the main result of this section.

Theorem 5.10. *Let $t \in M_{nn}$ be a doubly-F-astic matrix with $t_{ii} \in F$ for all $i = 1, \dots, n$, and W an arbitrary configuration function. Then t has a local weak decomposition if and only if $G(W)$ is strongly connected. Furthermore, there is at most one weak operation of \vee .*

We prove Theorem 5.10 in several steps. First we show that strong connectivity is a necessary condition. Then we derive a general decomposition method for a matrix t , and show how elementary matrices play a crucial role in determining that strong connectivity is sufficient. In the proofs we make no distinction between matrices with values in either belt, $R_{-\infty}$ or $R_{-\infty}^+$, and hence the operations of $+$ or $*$, denoted by the symbol \times . This is because we will use the isomorphisms to show the results hold for templates with values in $R_{-\infty}$ under the operation \boxtimes as well as for templates with values in $R_{-\infty}^+$ under the operation \oplus .

5.1.3. Necessary and Sufficient Conditions.

Our first theorem shows the sufficiency.

Theorem 5.11. *If every $t \in M_{nn}$ has a local decomposition with respect to W , then $G(W)$ is strongly connected.*

Proof: We assume by contradiction that $G(W)$ is not strongly connected, but every $t \in M_{nn}$ has a local decomposition with respect to W . Let $i, j \in V$ such that i is not reachable from j . Let

$$D(1) = \{ k \in V : i \text{ is reachable from } k \}$$

$$D(2) = \{ m \in V : m \text{ is reachable from } j \}.$$

Note that $D(1) \cap D(2) = \emptyset$, for otherwise $z \in D(1) \cap D(2)$ implies that i is reachable from z and z is reachable from j , giving i is reachable from j . Let t be a matrix which is local with respect to W .

First, we show that $t_{km} = -\infty$ if $k \in D(1)$ and $m \in D(2)$. Suppose by way of contradiction that $k \in D(1)$ and $m \in D(2)$ with $t_{km} \neq -\infty$. Then since t is local with respect to W , $m \in S_{-\infty}(t_k)$ implies that $m \in W(k)$. Thus, (m, k) is an edge and hence k is reachable from m . However, i is reachable from k , so i is reachable from m also. Thus, $m \in D(1)$, which is a contradiction, as m is also in $D(2)$. Thus, $t_{km} = -\infty$ if $k \in D(1)$ and $m \in D(2)$.

We now give a matrix which cannot have a local decomposition with respect to W .

Define $t \in M_{nn}$ and $v \in F_{-\infty}^n$ by

$$t_{km} = \begin{cases} \phi & \text{if } (k, m) = (i, j) \\ -\infty & \text{otherwise} \end{cases}, \quad v_k = \begin{cases} \phi & \text{if } k = j \\ -\infty & \text{otherwise} \end{cases}.$$

By assumption, there exists a decomposition of t , $t = \sum_{h=1}^r t^h$. At location k , the vector $b = t \times v$ has value

$$b_k = \sum_{h=1}^n t_{kh} \times v_h = t_{kj} \times v_j = \begin{cases} \phi & \text{if } k = i \\ -\infty & \text{if } k \neq i \end{cases}$$

Let $p \in \{1, \dots, r\}$. We show by induction on p that for $c(p) = (\sum_{h=1}^p t^h) \times v$, we have $c_u(p) = -\infty$ for every index $u \in D(1)$. If this is not true for $p = 1$, then there exists an index $u \in D(1)$ such that $c_u(1) \in F$, as $\sum_{k=1}^n t_{uk}^1 \times v_k = c_u(1)$. (Note that any t^k cannot have any $+\infty$ values, as then t would have $+\infty$ values.) Thus, there must exist a pair (u, m) such that $t_{um}^1 \in F$ and $v_m \in F$, which implies that $m = j$, which implies that $m \in D(2)$. However, $m \in D(2)$ implies that $t_{um}^1 = -\infty$ (as $u \in D(1)$), which is a contradiction. Thus, the claim must be true for $c(1)$. Now for some $p \in \{2, \dots, r\}$, by induction assume the claim is true for $c(1), \dots, c(p-1)$, and that it is not true for $c(p)$. Then there exists a $u \in D(1)$ such that $c_u(p) \in F$. Therefore there must exist an index m such that $t_{um}^p \in F$ and $c_m(p-1) \in F$. By the induction hypothesis, $m \notin D(1)$. Since t^p is local, we know that $m \in S_{-\infty}(t_u) \subset W(u)$ implies that u is reachable from m . Also, since $u \in D(1)$, i is reachable from u so $m \in D(1)$, which is a contradiction to the induction hypothesis. Therefore the claim that for $c(p) = (\sum_{h=1}^p t^h) \times v$, we have $c_u(p) = -\infty$ for every $u \in D(1)$ is true. In particular, it must be true for $c = c(r) = t \times v$. Thus, $c_i(p) = b_i = -\infty$, which contradicts our construction of c . Hence the proof is complete.

Q.E.D.

It can be easily shown that the set of all strictly doubly ϕ -astic $n \times n$ matrices with entries in $\{-\infty, \phi\}$ is isomorphic to the symmetric group S_n [39]. Since every permutation σ can be factored into a product of transpositions, every permutation matrix can be factored into a product of matrices corresponding to transpositions. This leads to our next definition of *exchange matrices*.

The *exchange matrix* $p^{ij} \in M_{nn}$ is the matrix defined by

$$p_{km}^{ij} = \begin{cases} \phi & \text{if } k = m \text{ and } k \neq i, k \neq j \\ & \text{or if } (k,m) = (i,j) \text{ or } (k,m) = (j,i) \\ -\infty & \text{otherwise} \end{cases}$$

The matrix p^{ij} corresponds to the transposition $\sigma = (i,j) \in S_n$. Note that $p^{ij} \times a = b$, where

$$b_k = \begin{cases} a_i & \text{if } k = j \\ a_j & \text{if } k = i \\ a_k & \text{otherwise} \end{cases}, \text{ for } k = 1, \dots, n.$$

Lemma 5.12. Let $(i,j) \in V$, $i \neq j$, let i,j denote a transposition in S_n , and let W be a configuration function. Suppose there exists an $i-j$ path in $G(W)$,

$$i = k_0, k_1, \dots, k_m = j.$$

Then the exchange matrix p^{ij} can be written as

$$p^{ij} = p^{ik_1} \times p^{k_1 k_2} \times \dots \times p^{k_{m-1} j} \times p^{k_{m-2} k_{m-1}} \times \dots \times p^{k_1 k_2} \times p^{ik_1}.$$

Proof: Note that $(i,j) \in S_n$ can be written as

$$(i,j) = (ik_1) \circ (k_1 k_2) \circ \dots \circ (k_{m-1} j) \circ (k_{m-2} k_{m-1}) \circ \dots \circ (k_1 k_2) \circ (ik_1).$$

In the cyclic notation for the permutation $\sigma = (i,j)$, let $i \rightarrow j$ denote i goes to j .

Then we have, for the above permutation, $i \rightarrow k_1, k_1 \rightarrow k_2, \dots, k_{m-1} \rightarrow j$, which

implies that $i \rightarrow j$. We also have $k_1 \rightarrow i \rightarrow k_1; k_2 \rightarrow k_1 \rightarrow k_2; \dots; k_{m-1} \rightarrow k_{m-2} \rightarrow$

$k_{m-1}; j \rightarrow k_{m-1} \rightarrow k_{m-2} \rightarrow \dots \rightarrow k_1 \rightarrow i$, which implies that $k_i \rightarrow k_i$, $i \neq 0, m$, and $j \rightarrow i$. We now use the fact that S_n is isomorphic to the set of all strictly doubly stochastic matrices, where the permutation $\sigma = (i_1, i_2, \dots, i_m)$ corresponds to the permutation matrix $p^{i_1, i_2, \dots, i_m} = p^\sigma$ defined by

$$p_{hk}^\sigma = \begin{cases} \phi & \text{if } (h, k) = (i_r, i_{r+1}), r = 1, \dots, m-1 \\ & \text{or if } (h, k) = (i_m, i_1) \\ -\infty & \text{otherwise} \end{cases}$$

and the product of two permutations σ_1, σ_2 corresponds to the matrix product

$p^{\sigma_1} \times p^{\sigma_2}$. Thus, for $\sigma = (i, j) =$

$(i k_1) \circ (k_1 k_2) \dots \circ (k_{m-1} j) \circ (k_{m-2} k_{m-1}) \circ \dots \circ (k_1 k_2) \circ (i k_1)$, we have

$$p^{ij} = p^{ik_1} \times p^{k_1 k_2} \times \dots \times p^{k_{m-1} j} \times p^{k_{m-2} k_{m-1}} \times \dots \times p^{k_1 k_2} \times p^{ik_1}.$$

Q.E.D.

Lemma 5.13. *Let $i, j \in V$, $i \neq j$ and let p^{ij} be the exchange matrix associated with i, j .*

Assume there exists an i - j path

$$i = k_0, k_1, \dots, k_m = j.$$

Then the exchange matrix p^{ij} has a local decomposition with respect to W .

Proof: If $i \in W(j)$ then we are done. Otherwise, use Lemma 5.12 and write

$$p^{ij} = p^{ik_1} \times p^{k_1 k_2} \times \dots \times p^{k_{m-1} j} \times p^{k_{m-2} k_{m-1}} \times \dots \times p^{k_1 k_2} \times p^{ik_1}.$$

Since (k_h, k_{h+1}) is an edge of the graph $G(W)$ for all $h = 0, \dots, m-1$, we know that

$k_h \in W(k_{h+1})$ and also that $k_{h+1} \in W(k_h)$ (as $G(W)$ is a graph), for all $h = 0, \dots, m-1$.

Also,

$$S_{-\infty}(p_r^{k_h k_{h+1}}) = \begin{cases} \{k_{h+1}\} & \text{if } r = h \\ \{k_h\} & \text{if } r = h+1 \\ \{r\} & \text{otherwise} \end{cases}$$

and hence $S_{-\infty}(p_r^{k_h, k_{h+1}}) \subset W(r)$ for all $r = 1, \dots, n$ and for all $h = 0, \dots, m-1$. Thus p^{ij} is local with respect to W for all $h = 0, \dots, m-1$ and hence

$$p^{ij} = p^{ik_1} \times p^{k_1 k_2} \times \dots \times p^{k_{m-1} j} \times p^{k_{m-2} k_{m-1}} \times \dots \times p^{k_1 k_2} \times p^{ik_1}$$

is a local decomposition of p^{ij} with respect to W .

Q.E.D.

Exchange matrices play an important role in the decomposition method, as we shall see shortly. We now present a decomposition method for lower triangular matrices.

The matrices ${}^k c$ as described in Lemma 5.8 are fairly sparse, but certainly not local with respect to most architectures. Lemma 5.9 showed that each matrix ${}^k c$ can be written as a product of even sparser matrices, and in the next theorem we show that these very sparse matrices can be decomposed locally. The matrices ${}^i c$ ($i > k$) are off matrices, and each is equivalent to a matrix which is local with respect to W , where $G(W)$ is strongly connected.

Theorem 5.14. *Let $b \in M_{nn}$ be a lower triangular off matrix with off-diagonal entry β at location (i, j) . Let W be a configuration function such that $G(W)$ is strongly connected. Then b is equivalent to a matrix which is local with respect to W . Furthermore, b has a local decomposition with respect to W .*

Proof: If $j \in W(i)$ then b is already local with respect to W . Otherwise, $W(i) \supseteq \{i\}$, and then we can find $k \in V$ such that $k \in W(i)$ ($k \neq i$), assuming without loss of generality that $k > i$. Let $q^{kj} = p^{kj}$. Then the matrix s defined by

$$s = p^{kj} \times b \times q^{kj}$$

is a matrix which is local with respect to W . To see this, note that $b \times q^{kj}$ exchanges columns j and k of matrix b . Thus,

$$(b \times q^{kj})_{hm} = \begin{cases} \phi & \text{if } h = m, (h,m) \neq (j,j) \text{ or } (h,m) \neq (k,k) \\ & \text{or if } (h,m) = (j,k) \text{ or } (h,m) = (k,j) \\ \beta & \text{if } (h,m) = (i,k) \\ -\infty & \text{otherwise} \end{cases}$$

The matrix $s = p^{kj} \times (b \times q^{kj})$ exchanges rows j and k of matrix $b \times q^{kj}$, so

$$(p^{kj} \times (b \times q^{kj}))_{hm} = \begin{cases} \phi & \text{if } h = m \\ \beta & \text{if } (h,m) = (i,k) \\ -\infty & \text{otherwise} \end{cases}$$

Note that $S_{-\infty}(s_i) = \{i,k\}$, and $S_{-\infty}(s_h) = \{h\}$ for $h \neq i$. Since $\{i,k\} \subset W(i)$, we have $S_{-\infty}(s_i) \subset W(i)$ and hence s is local with respect to W . Note also that s (in this case) is no longer lower triangular, as $k > i$.

A similar proof holds for the case $k < i$.

To show that b has a local decomposition, we write

$$b = p^{kj} \times s \times q^{kj},$$

as $(p^{h,m})^{-1} = p^{h,m}$ for any $h,m \in V$. Since $G(W)$ is strongly connected, there exists a j - k path for all pairs $(j,k) \in V$, and hence Lemma 5.13 applies to p^{jk} . Thus p^{jk} has a local decomposition, and since s is local and $q^{jk} = p^{jk}$, we have a local decomposition for b .

Q.E.D.

We are now in a position to prove the following general theorem.

Theorem 5.15. *Let $l \in M_{nn}$ be a lower triangular matrix with $l_{ii} \neq -\infty$ for all $i=1,\dots,n$ and W a configuration function such that $G(W)$ is strongly connected. Then l has a local decomposition with respect to W .*

Proof: By Corollary 5.6, l is equivalent to a matrix s where $c_{ii} = \phi$ for all i :

$$l = d \times c \times e = d \times c$$

where d is a diagonal matrix. By Lemma 5.8, c can be written as

$$c = {}^1c \times {}^2c \times \dots \times {}^{n-1}c.$$

By Lemma 5.9, each ${}^k c$ can be written as

$${}^k c = {}^{n,k}c \times {}^{n-1,k}c \times \dots \times {}^{k+1,k}c$$

where the ${}^{i,k}c$ are as in the statement of Lemma 5.9. Since each ${}^{i,k}c$ is an off matrix, we have, by Theorem 5.14, a local decomposition with respect to W for each ${}^{i,k}c$:

$${}^{i,k}c = \bigtimes_{j=1}^{m(i,k)} s(i,k,j)$$

where the $s(i,k,j)$ are the factors of the decomposition, and each $s(i,k,j)$ is local with respect to W , and the number of factors $m(i,k)$ is dependent on the values i and k .

Thus,

$$\begin{aligned} l &= d \times c = d \times [{}^1c \times {}^2c \times \dots \times {}^{n-1}c] \\ &= d \times [{}^{n1}c \times \dots \times {}^{21}c] \times \\ &\quad [{}^{n2}c \times \dots \times {}^{32}c] \times \dots \times {}^{n,n-1}c \\ &= d \times \left\{ \bigtimes_{j=1}^{m(n,1)} s(n,1,j) \times \dots \times \bigtimes_{j=1}^{m(2,1)} s(2,1,j) \right\} \times \dots \times \bigtimes_{j=1}^{m(n,n-1)} s(n,n-1,j) \end{aligned}$$

where $s(i,k,j)$ is local with respect to W for all i,k,j .

Q.E.D.

Using the property that $(s \times t)' = t' \times s'$ as stated in Theorem 5.1, we can prove the same sequence of theorems for an upper triangular matrix u satisfying $u_{ii} \neq -\infty$ for all $i = 1, \dots, n$. Thus we have

Theorem 5.16. *Let $u \in M_{nn}$ be an upper triangular matrix with $u_{ii} \in F$ for all i , and W a configuration function such that $G(W)$ is strongly connected. Then u has a local decomposition with respect to W .*

This leads immediately to the main theorem of this chapter.

Theorem 5.10. *Let $t \in M_{nn}$ be a doubly-F-astic matrix with $t_{ii} \in F$ for all $i = 1, \dots, n$, and let W be an arbitrary configuration function. Then t has a weak local decomposition with respect to W if and only if $G(W)$ is strongly connected. Furthermore, there is at most one weak operation of \vee .*

Proof: By Lemma 5.4, we know an arbitrary matrix t can be written in form

$$t = d \times s,$$

where $s_{ii} = \phi$ for all i and d is a diagonal matrix. By Lemma 5.5, $s = l \vee u$ where l and u are lower and upper triangular matrices, respectively, with $l_{ii} = u_{ii} = \phi$ for all i . Thus,

$$t = d \times (l \vee u).$$

Suppose that $G(W)$ is strongly connected. By Theorem 5.15, l has a local decomposition $l = \bigtimes_{j=1}^k c(j)$, and by Theorem 5.16, u has a local decomposition $u = \bigtimes_{i=1}^m r(i)$.

Hence, the expression

$$d \times \left\{ \left[\bigtimes_{j=1}^k c(j) \right] \vee \left[\bigtimes_{i=1}^m r(i) \right] \right\}$$

is a weak decomposition of $d \times [l \vee u]$, and, since each $c(j)$ and $r(i)$ is local with respect to W , $t = d \times [l \vee u] =$

$$d \times \left\{ \left[\bigtimes_{j=1}^k c(j) \right] \vee \left[\bigtimes_{i=1}^m r(i) \right] \right\}$$

is a weak local decomposition of t .

Theorem 5.11 shows the sufficiency of the statement.

Q.E.D.

As mentioned earlier, this is by no means the most efficient method of decomposition, and that one would not implement the decomposition by following the constructive steps used to prove Theorem 5.10. The results and hypothesis of the theorem are analogous to conditions for existence of solutions of differential equations. Thus, although the results guarantee the existence of local decompositions, efficient methods for computing them must still be developed.

5.2. Decomposition of Templates

We now present the results without proof of the previous section in context of the image algebra, using the isomorphism Ψ as defined in Chapter 2. As before, we let $V = \{1, \dots, n\}$ and W be a configuration function on V . We assume that $X \subset Z^2$ is finite and rectangular, of size $h \times k$, $hk = n$, with the lexicographical ordering as set out in Chapter 2, and that $F_{-\infty}$ is a sub-bounded l-group of $R_{-\infty}$ or $R_{-\infty}^+$. Everything in section 5.1 applies to templates under either \boxtimes or \odot . We will state the results using the symbol \boxtimes with the understanding that in this section, \boxtimes may be replaced everywhere by \odot and $+$ replaced everywhere by $*$, and the results will still be valid.

Let $a \in F_{\pm\infty}^Y$ and $b \in F_{\pm\infty}^X$ be arbitrary. We define the *outer product* of a and b to be an element $t \in (F_{\pm\infty}^X)^Y$, with gray values of

$$t_y(x) = a(y) + b(x).$$

We denote the outer product of a and b by $> a, b <$.

Let $t \in (F_{-\infty}^X)^X$. A *template decomposition* of t is a set of templates $t(1), \dots, t(j)$ such that $t = t(1) \boxtimes t(2) \boxtimes \dots \boxtimes t(j)$. The $t(i)$ are called the *factors* of the decomposition.

We write $t = \boxtimes_{i=1}^j t(i)$ is a decomposition of t . The decomposition is called *weak* if the operation \vee replaces any operation \boxtimes in the decomposition. We say $t \in (F_{-\infty}^X)^X$ is *local with respect to* W if $S_{-\infty}(t_{x_i}) \subset \{y_j : j \in W(i)\}$ for all $x_i \in X$. A decomposition $\{t(i)\}_{i=1}^j$ of

$t \in (F_{-\infty}^X)^X$ is called a *local decomposition of t with respect to W* if $t(i)$ is local with respect to W for all $i = 1, \dots, j$.

Lemma 5.17. *Let $s \in (F_{-\infty}^X)^W$, $t \in (F_{-\infty}^W)^Y$ be given. Then $(s \boxtimes t)' = t' \boxtimes s'$; the dual operations also satisfy this property.*

Lemma 5.18. *Suppose that $s \boxtimes t = r$ is a decomposition of r . Then this decomposition is not unique, and we have $\hat{s} \boxtimes \hat{t}$ is also a decomposition of r where*

$$\hat{s} = s \boxtimes \lambda, \quad \hat{t} = -\lambda \boxtimes t$$

and $\lambda \in F$ is arbitrary, $\lambda \in (F_{\pm\infty}^W)^W$.

Lemma 5.19. *Let $t \in (F_{-\infty}^X)^X$ be such that $t_x(x) \in F \forall x \in X$. Then t is equivalent to a template s which has the property that $s_x(x) = \phi \forall x \in X$. In this case, we have*

$$s = d \boxtimes t$$

where

$$d = \text{diag}(-t_{x_1}(x_1), -t_{x_2}(x_2), \dots, -t_{x_n}(x_n)).$$

If $s \in (F_{-\infty}^X)^X$ satisfies $s_x(x) = \phi$ for all $x \in X$, then we say that s is ϕ -diagonal. A template $t \in (F_{-\infty}^X)^X$ is said to be *lower diagonal* if $\Psi(t)$ is a lower diagonal matrix, and *upper diagonal* if $\Psi(t)$ is an upper diagonal matrix. If t is lower diagonal, then t satisfies

$$t_{x_j}(x_i) = -\infty \text{ if } i < j$$

and if t is upper diagonal then

$$t_{x_j}(x_i) = -\infty \text{ if } j < i.$$

Lemma 5.20. *Let $t \in (F_{-\infty}^X)^X$ be ϕ -diagonal. Then t has a weak decomposition into lower and upper triangular templates. In particular, t can be written as*

$$t = l \vee u,$$

where l (u) is lower (upper) diagonal, and $l_{x_i}(x_i) = u_{x_i}(x_i) = \phi$. Here,

$$l_{x_j}(x_i) = \begin{cases} t_{x_j}(x_i) & i \leq j \\ -\infty & \text{otherwise} \end{cases},$$

$$u_{x_j}(x_i) = \begin{cases} t_{x_j}(x_i) & j \leq i \\ -\infty & \text{otherwise} \end{cases}.$$

Corollary 5.21. *Let $t \in (F_{-\infty}^X)^X$ be lower or upper triangular with the property that $t_x(x) \in F \forall x \in X$. Then t is equivalent to a template which is ϕ -diagonal.*

A template $t \in (F_{-\infty}^X)^X$ is called an *off* template if $\Psi(t)$ is an off matrix. The *off-entry* value occurring at location (i,j) is the value $t_{x_j}(x_i)$.

Lemma 5.22. *If $l \in (F_{-\infty}^X)^X$ is lower triangular and ϕ -diagonal. Then*

$$l = {}^1r \boxtimes {}^2r \boxtimes \dots \boxtimes {}^{n-1}r,$$

where

$${}^k r = 1 \vee [> l_{x_k}, l_{x_k} <],$$

and $>, <$ denotes the outer product.

Note that $s = > l_{x_k}, l_{x_k} < \in (F_{-\infty}^X)^X$ and

$$s_{x_h}(x_m) = l_{x_k}(x_h) + l_{x_k}(x_m) = \begin{cases} l_{x_k}(x_m) & h = k \\ -\infty & h \neq k \end{cases}.$$

Thus, $s_{x_h} = -\infty$ (the null image on X), if $h \neq k$, and $s_{x_h} = l_{x_k}$ otherwise. Of course, we have $\Psi(s) = {}^k c$ in this case.

Lemma 5.23. *Let $l \in (F_{-\infty}^X)^X$ be lower triangular and ϕ -diagonal, and let ${}^k r$ be as in*

Lemma 5.22, $k=1, \dots, n$. Then

$${}^k r = {}^{n,k} r \boxtimes {}^{n-1,k} r \boxtimes \dots \boxtimes {}^{k+1,k} r,$$

where ${}^{i,k} r \in (F_{-\infty}^X)^X$ is defined to be

$${}^{i,k} r_y(x) = \begin{cases} \phi & y = x \\ l_{x_k}(x_i) & y = x_k, x = x_i \\ -\infty & \text{otherwise} \end{cases}$$

for $i = k+1, \dots, n$.

Note that each template ${}^{i,k} r$ is an off template, with off-entry value occurring at location (i, k) .

The main result of this chapter is

Theorem 5.24. *Let $t \in (F_{-\infty}^X)^X$ be a doubly-F-astic template with $t_x(x) \in F$ for all $x \in X$, and W a configuration function. Then t has a weak local decomposition if and only if $G(W)$ is strongly connected. Furthermore, there is at most one weak operation of \vee .*

The sufficiency of Theorem 5.24 is shown next in Theorem 5.25.

Theorem 5.25. *If every $t \in (F_{-\infty}^X)^X$ has a local decomposition with respect to W , then $G(W)$ is strongly connected.*

The exchange template $p^{x_i x_j} \in (F_{-\infty}^X)^X$ associated with (i, j) is the template defined by

$$p_z^{x_i x_j}(u) = \begin{cases} \phi & \text{if } z = u \text{ and } z \neq x_i, z \neq x_j \\ & \text{or if } (u, z) = (x_i, x_j) \text{ or } (u, z) = (x_j, x_i) . \\ -\infty & \text{otherwise} \end{cases}$$

The template $p^{x_i x_j}$ corresponds to the transposition $\sigma = (i, j) \in S_n$ and to the exchange matrix p^{ij} . Note that $a \boxtimes p^{x_i x_j} = b$, where

$$b(z) = \begin{cases} a(x_i) & \text{if } z = x_j \\ a(x_j) & \text{if } z = x_i \\ a(x_k) & \text{otherwise} \end{cases} .$$

Lemma 5.26. *Let $x_i, x_j \in X$, $i \neq j$, and let W be a configuration function. Suppose there exists an $i - j$ path in $G(W)$,*

$$i = k_0, k_1, \dots, k_m = j .$$

Then the exchange template $p^{x_i x_j}$ can be written as

$$p^{x_i x_j} = p^{x_i x_{k_1}} \boxtimes p^{x_{k_1} x_{k_2}} \boxtimes \dots \boxtimes p^{x_{k_{m-1}} x_j} \boxtimes p^{x_{k_{m-2}} x_{k_{m-1}}} \boxtimes \dots \boxtimes p^{x_{k_1} x_{k_2}} \boxtimes p^{x_i x_{k_1}} .$$

Lemma 5.27. *Let $x_i, x_j \in X$, $i \neq j$, and let W be a configuration function. Let $p^{x_i x_j}$ be the exchange template associated with i, j . Assume there exists an $i - j$ path*

$$i = k_0, k_1, \dots, k_m = j .$$

Then the exchange template $p^{x_i x_j}$ has a local decomposition with respect to W .

We now present a local decomposition method for an off template.

Theorem 5.28. *Let $s \in (F_{-\infty}^X)^X$ be a lower triangular, ϕ -diagonal template with off-entry value of β at location (i, j) :*

$$s_{x_k}(x_m) = \begin{cases} \phi & k = m \\ \beta & (k, m) = (j, i) . \\ -\infty & \text{otherwise} \end{cases}$$

Let W be configuration function such that $G(W)$ is strongly connected. Then s is equivalent to a template which is local with respect to W . Furthermore, s has a local decomposition with respect to W .

Theorem 5.29. Let $l \in (F_{-\infty}^X)^X$ be a lower triangular template with $l_x(x) \in F$ for all $x \in X$, and W a configuration function such that $G(W)$ is strongly connected. Then l has a local decomposition with respect to W .

Using the property of transpose as stated in Theorem 5.17, we can prove theorems 5.19, 5.21, 5.22, 5.23, 5.28 and 5.29 for upper triangular templates.

Theorem 5.30. Let $u \in (F_{-\infty}^X)^X$ be an upper triangular template with $u_{x_i}(x_i) \in F$ for all i , and W a configuration function such that $G(W)$ is strongly connected. Then u has a local decomposition with respect to W .

The main theorem follows immediately.

Theorem 5.24. Let $t \in (F_{-\infty}^X)^X$ be a doubly-F-astic template with $t_x(x) \in F$ for all $x \in X$, and W a configuration function. Then t has a weak local decomposition if and only if $G(W)$ is strongly connected. Furthermore, there is at most one weak operation of \vee .

The use of this theorem is made clear by the following discussion. Suppose a lattice transform t is to be mapped to a parallel architecture through a decomposition technique, whereby

$$t = \bigvee_{i=1}^h r_i \vee \bigvee_{j=1}^k s_j.$$

Applying the transform t to an image a , we have

$$a \boxtimes t = a \boxtimes \left[\bigvee_{i=1}^h r_i \vee \bigvee_{j=1}^k s_j \right] = \left[a \boxtimes \left[\bigvee_{i=1}^h r_i \right] \right] \vee \left[a \boxtimes \left[\bigvee_{j=1}^k s_j \right] \right]$$

$$= \left[(\cdots ((a \boxtimes r_1) \boxtimes r_2) \cdots) \boxtimes r_h \right] \vee \left[(\cdots ((a \boxtimes s_1) \boxtimes s_2) \cdots) \boxtimes s_k \right].$$

Every r_i, s_j is local with respect to the network of processors and hence directly implementable on the parallel architecture. The one operation of maximum is related to storage.

When computing the transform $a \boxtimes t$, the image $\left[(\cdots ((a \boxtimes r_1) \boxtimes r_2) \cdots) \boxtimes r_h \right]$ must be computed, the result stored, the image $\left[(\cdots ((a \boxtimes s_1) \boxtimes s_2) \cdots) \boxtimes s_k \right]$ computed, and then the maximum between the two results taken. Most processing elements in parallel architectures have a small amount of local memory available. If they do not, then the host machine, or another computer which is able to communicate with the parallel one, must store the results separately then return both to the parallel machine to compute the last pointwise maximum. Thus the one operation of maximum is not an unreasonable stipulation in a decomposition.

5.3. Applications to Rectangular Templates

A special class of templates within the invariant ones are *rectangular* templates. Let $X \subset Z^2, |X| = m, |Y| = n$. Then a *rectangular template* $t \in (F^X)^X$ has a rectangular array for its support (away from the boundary of X). Specifically, $S_{-\infty}(t_y)$ is a rectangle, for some $y \in X$ such that $S_{-\infty}(t_y) \cap \delta X = \emptyset$, where $\delta X = \{(i,j) : i=0 \text{ or } m-1, \text{ or } j=0 \text{ or } n-1\}$. We will use the fact that invariant templates correspond to matrices that are block toeplitz with toeplitz blocks [33]. First, we derive the results in matrix notation, and use the isomorphism to map the theorems to image algebra notation. We give conditions under which rectangular templates can be decomposed into the product of two strip templates, one vertical and one horizontal. Since the proof of this theorem is constructive, a decomposition is given. This theorem was presented in its original form by Li [61].

5.3.1. Decomposition of block toeplitz matrices

This section presents results on a decomposition technique for block toeplitz matrices with toeplitz blocks. Matrices that are block toeplitz with toeplitz blocks correspond to invariant templates. Conditions are given to guarantee a decomposition of such matrices into a product of two matrices, each of which correspond to a *strip* template in the image algebra. Strip templates are templates whose supports are a $1 \times k$ array (horizontal), or a $k \times 1$ array (vertical). Even though these decompositions may not be local ones, a decomposition of this type will reduce the number of computations per pixel necessary to compute the transform, on either a parallel or sequential machine. As an example, suppose $r \in (F^X)^X$ is a rectangular template and has a $h \times k$ support. Suppose there exists two templates t and s such that $r = s \boxtimes t$, where t is a rectangular template with a $h \times 1$ support and s is a rectangular template with a $1 \times k$ support. Then for $a \in F^X$ and using the associative property of the \boxtimes operation, we have

$$a \boxtimes r = a \boxtimes (s \boxtimes t) = (a \boxtimes s) \boxtimes t.$$

Computing $a \oplus r$ directly involves finding the maximum of hk additions, while computing $(a \boxtimes s) \boxtimes t$ involves finding the maximum of $h + k$ additions. Thus savings in computation can be realized on sequential as well as parallel processors.

Minimax matrix results. We shall first prove the following lemma.

Lemma 5.31. *Suppose that $s \in M_{nn}$ is a diagonal matrix of form $s = \text{diag}(\alpha, \dots, \alpha)$, where $\alpha \in F$. Then for $t \in M_{nn}$, $s \times t = t \times s$.*

Proof: Let $r = s \times t$, and $w = t \times s$. Then

$$r_{ij} = \bigvee_{k=1}^n s_{ik} \times t_{kj} = s_{ij} \times t_{jj}$$

and

$$w_{ij} = \bigvee_{k=1}^n t_{ik} \times s_{kj} = t_{ii} \times s_{ij}.$$

Since $t_{ii} = t_{jj} = \alpha \forall i, j = 1, \dots, n$, we have

$$r_{ij} = s_{ij} \times t_{jj} = s_{ij} \times t_{ii} = t_{ii} \times s_{ij} = w_{ij}.$$

Q.E.D.

Lemma 5.32. Suppose $s, t \in M_{nn}$ are each block toeplitz with toeplitz blocks. Denote s and t by

$$s = \begin{bmatrix} \sigma^{11} & \dots & \sigma^{1h} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \sigma^{h1} & \dots & \sigma^{hh} \end{bmatrix} \quad t = \begin{bmatrix} \tau^{11} & \dots & \tau^{1h} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \tau^{h1} & \dots & \tau^{hh} \end{bmatrix}.$$

Suppose s and t satisfy the following conditions:

(1) $n = hk$, and each block σ^{ij}, τ^{ij} is a $k \times k$ matrix;

(2) there exist $2h-1$ constants λ_{ij} such that $\tau^{ij} = \text{diag}(\lambda_{ij}, \dots, \lambda_{ij})$ and the matrix

$\Lambda = (\lambda_{ij}) \in M_{hh}$ is toeplitz;

(3) $\sigma^{ij} = \begin{cases} \Phi_k & i \neq j \\ a \in M_{hh} & i = j \end{cases}$ where Φ_k denotes the null matrix of size $k \times k$, and a is a toeplitz matrix where $a_{ii} \in F$;

If (1) - (3) are satisfied, then $s \times t = t \times s$ and is of form

$$\begin{bmatrix} \sigma^{11} \times \tau^{11} & \sigma^{11} \times \tau^{12} & \dots & \sigma^{11} \times \tau^{1h} \\ \sigma^{22} \times \tau^{21} & \sigma^{22} \times \tau^{22} & \dots & \sigma^{22} \times \tau^{2h} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma^{hh} \times \tau^{h1} & \cdot & \dots & \sigma^{hh} \times \tau^{hh} \end{bmatrix}.$$

Proof: Let $u = s \times t$, and denote the matrix $u = (u_{ij})$ in block notation by

$$u = \begin{bmatrix} \mu^{11} & . & . & . & \mu^{1h} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \mu^{h1} & . & . & . & \mu^{hh} \end{bmatrix}.$$

Then

$$u_{ij} = \bigvee_{m=1}^n s_{im} + t_{mk}.$$

For $i, k \in Z^+$ and using the division algorithm for integers, we have

$$i = k \cdot p_i + r_i \quad \text{for some } 0 \leq p_i \leq h \text{ and } 0 \leq r_i \leq k-1.$$

If $r_i = 0$ then write $i = k \cdot (p_i - 1) + k$, where $0 \leq p_i \leq h-1$. Thus, (abusing notation) in this case, we can always write

$$i = k \cdot p_i + r_i \quad \text{for some } 0 \leq p_i \leq h-1, \text{ and } 0 < r_i \leq k.$$

Similarly, we have

$$j = k \cdot p_j + r_j \quad \text{for some } 0 \leq p_j \leq h-1, \text{ and } 0 < r_j \leq k.$$

The element u_{ij} lies in the (p_i+1, p_j+1) -th block of u , and is the (r_i, r_j) -th entry in that block. The matrix s has values of $-\infty$ when m indexes outside the block σ^{p_i+1, p_i+1} , i.e.,

$$s_{im} = -\infty \text{ for values of } m \text{ not in } \{w : i - r_i + 1 \leq w \leq i - r_i + k\}.$$

Thus, $u_{ij} = \bigvee_{m=i-r_i+1}^{i-r_i+k} s_{im} + t_{mj}$. Since the values s_{im} , $i - r_i + 1 \leq m \leq i - r_i + k$, lie on

the r_i -th row of block σ^{p_i+1, p_i+1} , and the values t_{mj} lie on the r_j -th column of block τ^{p_j+1, p_j+1} , the value u_{ij} can be represented by

$$u_{ij} = (\sigma^{p_i+1, p_i+1})_{r_i} \times (\tau^{p_j+1, p_j+1})_{r_j}.$$

This is true for any value of u_{ij} such that u_{ij} lies in block μ^{p_i+1, p_j+1} , i.e.,

$$(\sigma^{p_i+1, p_i+1})_{r_i} \times (\tau^{p_i+1, p_j+1})^{r_j} = (\mu^{p_i+1, p_j+1})_{r_i, r_j}, \quad 1 \leq r_i, r_j \leq k.$$

Thus we have

$$\sigma^{p_i+1, p_i+1} \times \tau^{p_i+1, p_j+1} = \mu^{p_i+1, p_j+1}.$$

Since i, j were arbitrary, this relation holds for each value $1 \leq p_i+1, p_j+1 \leq h$.

That is,

$$\sigma^{ij} \times \tau^{ij} = \mu^{ij}, \quad \forall i, j = 1, \dots, n.$$

To show commutativity of the product, we note that for all i, j , σ^{ii} is toeplitz with finite elements on the diagonal, and the block τ^{ij} is a diagonal matrix with a constant value on its diagonal. Thus, we have by Lemma 5.31, for all i, j ,

$$\sigma^{ii} \times \tau^{ij} = \tau^{ij} \times \sigma^{ii}.$$

Let $w = t \times s$, $w = (\omega^{ij})$ in block notation, $i, j = 1, \dots, h$. We show that

$$\omega^{ij} = \tau^{ij} \times \sigma^{ij}. \quad \text{Here, } w_{ij} = \bigvee_{m=1}^n t_{im} \times s_{mj}. \quad \text{As before and using the same method and}$$

notation, we have the same restrictions on the indices of s_{mj} :

$$s_{mj} = -\infty \text{ for values of } m \text{ not in } \{w : j - r_j + 1 \leq w \leq j - r_j + k\},$$

and the values s_{mj} for $j - r_j + 1 \leq m \leq j - r_j + k$ lie on the r_j -th column of the block σ^{p_j+1, p_j+1} . Similarly, the values t_{im} for $j - r_j + 1 \leq m \leq j - r_j + k$ lie on the r_i -th row of block τ^{p_i+1, p_j+1} , i.e.,

$$(\tau^{p_i+1, p_j+1})_{r_i} \times (\sigma^{p_i+1, p_j+1})^{r_j} = (\omega^{p_i+1, p_j+1})_{r_i, r_j},$$

and thus in general we have

$$\tau^{ij} \times \sigma^{ij} = \omega^{ij}.$$

Since $\sigma^{ii} = \sigma^{jj}$ for all i, j , we have

$$\omega^{ij} = \tau^{ij} \times \sigma^{ij} = \tau^{ij} \times \sigma^{ii} = \sigma^{ii} \times \tau^{ij} = \mu^{ij}.$$

Q.E.D.

Let us define a *strip matrix* $s \in M_{nn}$ if $\Psi^{-1}(s)$ is a strip template. A *vertical strip matrix* $t \in M_{nn}$, $n = hk$, is a matrix which is block toeplitz with toeplitz blocks, $t = (\tau^{ij})$, each τ^{ij} is $k \times k$, $i, j = 1, \dots, h$, and $\tau^{ij} = \text{diag}(\alpha_{ij}, \dots, \alpha_{ij})$ for some $\alpha_{ij} \in F_{-\infty}$. Since $\tau^{ij} = \tau^{i-j+1, 1}$ if $j \geq i$ and $\tau^{ij} = \tau^{i-j+1, 1}$ if $i \geq j$, there are actually only $2h-1$ constants α_{ij} which determine the τ^{ij} . Denote these constants by $\alpha_{11}, \alpha_{i1}, \alpha_{ij}$, $i, j = 2, \dots, h$, where

$$\tau^{ij} = \begin{cases} \text{diag}(\alpha_{11}, \dots, \alpha_{11}) & \text{if } i = j \\ \text{diag}(\alpha_{ij}, \dots, \alpha_{ij}) & \text{if } j > i \\ \text{diag}(\alpha_{i1}, \dots, \alpha_{i1}) & \text{if } i > j \end{cases}$$

Also, we must have $\alpha_{ij} \in F$ for $j = 1, 2, \dots, m_1$ and $\alpha_{i1} \in F$ for $j = 1, 2, \dots, m_2$, for some $1 \leq m_1, m_2 \leq h$. Vertical strip matrices correspond to vertical strip templates.

A *horizontal strip matrix* $s \in M_{nn}$, $n = hk$, is a matrix which is block toeplitz with toeplitz blocks, $s = (\sigma^{ij})$, each σ^{ij} is $k \times k$, $i, j = 1, \dots, h$, and $\sigma^{ij} = \begin{cases} \Phi_k & i \neq j \\ a \in M_{hh} & i = j \end{cases}$, where Φ_k denotes the null matrix of size $k \times k$, and a is a toeplitz matrix where $a_{ij} \in F$. Horizontal strip matrices correspond to horizontal strip templates. Note that such a matrix is a *diagonal* block toeplitz matrix with toeplitz blocks, that is, the only non-null block is the diagonal block.

We are now in a position to prove the main result of this section.

Theorem 5.33. Assume $u \in M_{nn}$, $n = hk$, is block toeplitz with toeplitz blocks, and write

$$u = \begin{bmatrix} \mu^{11} & \dots & \mu^{1h} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \mu^{h1} & \dots & \mu^{hh} \end{bmatrix}$$

where each μ^{ij} is a $k \times k$ submatrix and $(\mu^{ij})_{mm} \neq -\infty \forall i, m$. Then u is decomposable into two strip matrices, one horizontal s and one vertical t , if and only if

(1) \exists $2h-1$ constants λ_{ij} , λ_{i1} , $i, j = 1, \dots, h$ with $\lambda_{ii} = 0$, such that $\Lambda = (\lambda_{ij}) \in M_{hh}$ is toeplitz and of form

$$\Lambda = \begin{bmatrix} 0 & \lambda_{12} & \dots & \lambda_{1k} \\ \cdot & 0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \lambda_{k1} & \lambda_{k2} & \dots & 0 \end{bmatrix};$$

(2) the constants λ_{ij} satisfy $\lambda_{ij} \times \mu^{ii} = \mu^{ij}$ for all $i, j = 1, \dots, k$.

In this case a decomposition is given by the following matrices which are each block toeplitz with toeplitz blocks:

$$s = \begin{bmatrix} \sigma^{11} & \dots & \sigma^{1h} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \sigma^{h1} & \dots & \sigma^{hh} \end{bmatrix} \quad t = \begin{bmatrix} \tau^{11} & \dots & \tau^{1h} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \tau^{h1} & \dots & \tau^{hh} \end{bmatrix}$$

where each of σ^{ij} , τ^{ij} are $k \times k$ matrices and satisfy:

$$\tau^{ij} = \text{diag}(\lambda_{ij}, \dots, \lambda_{ij}); \quad \sigma^{ij} = \begin{cases} \Phi_k & i \neq j \\ \mu^{ii} \in M_{hh} & i = j \end{cases}$$

Here, $\lambda_{11} = 0$, and $\lambda_{ij} = u_{ij} \times (u_{11})^{-1}$, $\lambda_{i1} = u_{i1} \times (u_{11})^{-1}$, $i, j = 2, \dots, h$.

Proof: Suppose (1) and (2) are satisfied. We show that for s and t as in the statement of

the theorem, $s \times t = u$. Let $w = s \times t$. By Lemma 5.32, w has form

$$w = \begin{bmatrix} \sigma^{11} \times \tau^{11} & \dots & \sigma^{11} \times \tau^{1h} \\ \sigma^{22} \times \tau^{21} & \dots & \sigma^{22} \times \tau^{2h} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \sigma^{hh} \times \tau^{h1} & \dots & \sigma^{hh} \times \tau^{hh} \end{bmatrix}$$

Choose an element w_{ij} . As in Lemma 5.32,

$$i = k \cdot p_i + r_i \quad \text{for some } 0 \leq p_i \leq h-1, \text{ and } 0 < r_i \leq k.$$

$$j = k \cdot p_j + r_j \quad \text{for some } 0 \leq p_j \leq h-1, \text{ and } 0 < r_j \leq k.$$

and the pair (i, j) lies in the block $(p_i + 1, p_j + 1)$. Thus,

$$\begin{aligned} w_{ij} &= (s \times t)_{ij} = (\sigma^{p_i+1, p_i+1})_{r_i} \times (\tau^{p_j+1, p_j+1})_{r_j} \\ &= \bigvee_{m=i-r_i+1}^{i-r_i+k} u_{im} + t_{mj}. \end{aligned}$$

Since $\tau^{p_i+1, p_j+1} = \text{diag}(\lambda_{p_i+1, p_j+1}, \dots, \lambda_{p_i+1, p_j+1})$, the only possible non-null element of $\{t_{mj} : i-r_i+1 \leq i-r_i+k\}$ lies on the diagonal of τ^{p_i+1, p_j+1} , and, in fact, lies in the (r_j, r_j) -th position in τ^{p_i+1, p_j+1} . The (r_j, r_j) -th position in τ^{p_i+1, p_j+1} is in the $(i, i-r_i+r_j)$ -th position in u . So

$$w_{ij} = u_{i, i-r_i+r_j} \times \lambda_{p_i+1, p_j+1}.$$

By (1) we know that

$$\lambda_{p_i+1, p_j+1} \times \mu^{p_i+1, p_i+1} = \mu^{p_i+1, p_j+1}$$

which implies that

$$(\lambda_{p_i+1, p_j+1})_{r_i} \times (\mu^{p_i+1, p_i+1})_{r_j} = (\mu^{p_i+1, p_j+1})_{r_i, r_j} = u_{ij}$$

which implies

$$\lambda_{p_i+1, p_j+1} \times (\mu^{p_i+1, p_i+1})_{r_i, r_j} = u_{ij}.$$

But

$$(\mu^{p_i+1, p_j+1})_{r_i, r_j} = u_{i, i-r_i+r_j}.$$

Therefore,

$$u_{ij} = u_{i, i-r_i+r_j} \times \lambda_{p_i+1, p_j+1} = w_{ij}.$$

Now suppose that u is decomposable into two strip matrices, a horizontal one s and a vertical one t . We show that s and t satisfy conditions (1) - (3). By the definition of a vertical strip matrix, we see that t is block toeplitz with toeplitz blocks, and

$t = r^{ij}$ is of form

$$r^{ij} = \begin{cases} \text{diag}(\alpha_{11}, \dots, \alpha_{11}) & \text{if } i = j \\ \text{diag}(\alpha_{1j}, \dots, \alpha_{1j}) & \text{if } j > i \\ \text{diag}(\alpha_{i1}, \dots, \alpha_{i1}) & \text{if } i > j \end{cases}$$

for some $2h-1$ constants $\alpha_{11}, \alpha_{i1}, \alpha_{1j}, i, j = 2, \dots, h$. Set $\alpha_{11} = 0, \alpha_{1j} = u_{1j} \times (u_{11})^{-1}$,

$$\alpha_{i1} = u_{i1} \times (u_{11})^{-1}. \text{ Define } \Lambda = (\lambda_{ij}) \in M_{hh} \text{ by } \lambda_{ij} = \begin{cases} \alpha_{11} & \text{if } i = j \\ \alpha_{11} & \text{if } j < i \\ \alpha_{1j} & \text{if } i < j \end{cases} \text{ Then } \Lambda \text{ is}$$

toeplitz, with the $2h-1$ constants $\lambda_{11}, \lambda_{i1}, \lambda_{1j}, i, j = 2, \dots, h$, satisfying condition (1).

Also, s is a horizontal strip matrix, which means that (using block notation)

$s = (\sigma^{ij})$, where

$$\sigma^{ij} = \begin{cases} \Phi_k & i \neq j \\ a \in M_{hh} & i = j \end{cases}$$

and a is toeplitz with non-null diagonal entries. Set $a = \mu^{ii}$. The last thing we need to show is that condition (2) is satisfied. By Lemma 5.32, the (i,j) -th block of $s \times t$ is of form

$$\sigma^{ii} \times r^{ij},$$

and since r^{ij} is diagonal, $\sigma^{ii} \times r^{ij} = r^{ij} \times \sigma^{ii}$. So, $r^{ij} \times \sigma^{ii} = \text{diag}(\alpha_{ij}, \dots, \alpha_{ij}) \times \mu^{ii} = \lambda_{ij} \times \mu^{ii} = \mu^{ij}$. Obviously, $s \times t = u$, and this completes the proof.

Q.E.D.

The statement in the image algebra of Theorem 5.33 is

Theorem 5.34 [61]. *Let $r \in (R_{\pm\infty}^X)^X$ be a rectangular template with non-null weights $r_{x_i}(x_j)$, $i = 1, \dots, m$, $j = 1, \dots, k$. Then r has a decomposition into two strip templates, one horizontal and one vertical, if and only if for all $1 \leq i, i' \leq m$ and $1 \leq j, j' \leq k$,*

$$r_{x_i}(x_j) - r_{x_{i'}}(x_j) = r_{x_i}(x_{j'}) - r_{x_{i'}}(x_{j'}).$$

CHAPTER 6 THE DIVISION ALGORITHM

6.1. A Division Algorithm in a Non-Euclidean Domain

The integers have the property that a *division algorithm* can be defined on them. For $a, b \in \mathbb{Z}$, there exist unique integers q, r such that $a = qb + r$ where $|r| < |b|$. This is an example of an integral domain upon which is defined a *Euclidean valuation* [61]. In this section we present a division algorithm for the minimax algebra structure, and give an application of this result to image processing in the image algebra notation.

We remark that the boolean case has already been stated by P. Miller [62], and will be discussed in more detail at the end of this section.

6.1.1. A Matrix Division Algorithm

Let $E^1 = F_{-\infty}$ be a sub-bounded l-group of $R_{-\infty}$. For notational convenience, we will write $t \in M_{nn}(-\infty)$ when we mean that the matrix t will assume values only on $F \cup \{-\infty\}$. Similarly, we write $t \in M_{nn}(+\infty)$ when the matrix t assumes values only on $F \cup \{+\infty\}$. We will show that for a finite vector $a \in F^n$ and a subset of matrices of $M_{nn}(-\infty)$, that there exist vectors q and $r \in F^{n-\infty}$ such that

$$a = (t' \times q) \vee r$$

Lemma 6.1. *Let $a \in F^n$ be finite, and $t \in M_{nn}(-\infty)$ satisfy $S_{-\infty}(t_i) \neq \emptyset \forall i = 1, \dots, n$.*

Define \hat{t} by $\hat{t} \equiv (t^)'$. Then both $\hat{t} \times' a$ and $t' \times (\hat{t} \times' a)$ are finite. and*

$$t' \times (\hat{t} \times' a) \leq a.$$

Proof: First we note that

$$(\hat{t})_{ij} = [(t^*)']_{ij} = (t^*)_{ji} = (t_{ij})^* = \begin{cases} -t_{ij} & \text{if } t_{ij} \in F \\ +\infty & \text{if } t_{ij} = -\infty \end{cases}$$

and that $S_{+\infty}(\hat{t}_i) = S_{-\infty}(t_i)$. Let $b = \hat{t} \times' a$ and let $c = t' \times b$. At location i ,

$b_i = \bigwedge_{j \in S_{+\infty}(\hat{t}_i)} \hat{t}_{ij} + a_j$. The vector b is finite, as for $i \in \{1, 2, \dots, n\}$, there exists at

least one $j \in S_{+\infty}(\hat{t}_i)$ such that $\hat{t}_{ij} + a_j$ is finite, since by hypothesis, $S_{-\infty}(t_i) \neq \emptyset \forall i$

$= 1, \dots, n$, and $a_i \in F \forall i = 1, \dots, n$ also. Thus, $b_i = \left\{ \bigwedge_{j \in S_{+\infty}(\hat{t}_i)} \hat{t}_{ij} + a_j \right\} \in F \forall i$. At

location i , $c_i = \bigvee_{j \in S_{-\infty}(t_i)} t'_{ij} + b_j$. By a similar argument, we see that $c_i \in F \forall i$.

Suppose that $c_i = t'_{ik} + b_k$ for some $k \in \{1, 2, \dots, n\}$. Then $b_k = \bigwedge_{j \in S_{+\infty}(\hat{t}_k)} \hat{t}_{kj} + a_j =$

$\hat{t}_{kp} + a_p$ for some p . Since $k \in S_{-\infty}(t'_i)$, we know that $i \in S_{-\infty}(t_k)$. Since $S_{-\infty}(t_k) =$

$S_{+\infty}(\hat{t}_k)$, we know that $i \in S_{+\infty}(\hat{t}_k)$, and, hence

$$\hat{t}_{kp} + a_p \leq \hat{t}_{ki} + a_i \in F,$$

by our choice of p and the fact that $\hat{t}_{ki} \in F$ and $a_i \in F \forall i$. Thus

$$c_i = t'_{ik} + b_k = t'_{ik} + \hat{t}_{kp} + a_p \leq t'_{ik} + \hat{t}_{ki} + a_i = t_{ki} + (-t_{ki}) + a_i = a_i$$

Thus, $c_i \leq a_i$, and our lemma is proved.

Q.E.D.

We now state the Division Algorithm.

Theorem 6.2. The Division Algorithm. *Let a, t satisfy the hypothesis of Lemma 6.1.*

Then for $q = \hat{t} \times' a$, and r defined by

$$r_i = \begin{cases} a_i & \text{if } a_i > [t' \times (\hat{t} \times' a)]_i \\ -\infty & \text{if } a_i = [t' \times (\hat{t} \times' a)]_i \end{cases}$$

we have

$$a = (t' \times q) \vee r$$

Proof: By Lemma 6.1, $a \geq t' \times q = t' \times (\hat{t} \times' a)$, and hence,

$$a \geq r.$$

Thus, $[t' \times (\hat{t} \times' a)] \vee r \leq a$. To show that equality holds, that is, that

$[t' \times (\hat{t} \times' a)] \vee r = a$, we examine two cases.

Case 1. $a_i > [t' \times (\hat{t} \times' a)]_i$. Here, $[t' \times (\hat{t} \times' a)]_i \vee r_i = [t' \times (\hat{t} \times' a)]_i \vee a_i = a_i$.

Case 2. $a_i = [t' \times (\hat{t} \times' a)]_i$. Here, $[t' \times (\hat{t} \times' a)]_i \vee r_i = a_i \vee r_i = a_i \vee -\infty = a_i$.

Q.E.D.

Now suppose we have $a = (t' \times q) \vee r$ for a finite, $t \in M_{nn}(-\infty)$ and t satisfying

$S_{-\infty}(t_i) \neq \emptyset \forall i = 1, \dots, n$. Define

$$a^0 = a$$

$$r^0 = r, \text{ and}$$

$$a^i = \hat{t} \times' a^{i-1}.$$

Then we have

$$a = a^0 = (t' \times a^1) \vee r^0 \tag{6-1}$$

By Lemma 6.1, $a^1 = \hat{t} \times' a^0$ is finite, and, in fact, $a^i = \hat{t} \times' a^{i-1}$ will be finite for each $i = 1, 2, \dots$. Thus, the Division Algorithm applies in particular to a^1 :

$$a^1 = (t' \times a^2) \vee r^1 \tag{6-2}$$

and substituting (6-2) into (6-1), we get

$$a = (t' \times a^1) \vee r^0$$

$$\begin{aligned}
&= \{ t' \times [(t' \times a^2) \vee r^1] \} \vee r_0 \\
&= (t' \times t' \times a^2) \vee (t' \times r^1) \vee r^0 \\
&= [(t')^2 \times a^2] \vee (t' \times r^1) \vee r^0
\end{aligned} \tag{6-3}$$

where $(t')^k$ denotes the k -fold product of t , $\bigtimes_{i=1}^k (t')$.

Apply the Division Algorithm to a^2 , to get

$$a^2 = (t' \times a^3) \vee r^2$$

and substituting this into (6-3), we get

$$\begin{aligned}
a &= \{ (t')^2 \times [(t' \times a^3) \vee r^2] \} \vee (t' \times r^1) \vee r^0 \\
&= [(t')^2 \times t' \times a^3] \vee [(t')^2 \times r^2] \vee [t' \times r^1] \vee r^0 \\
&= [(t')^3 \times a^3] \vee [(t')^2 \times r^2] \vee [t' \times r^1] \vee r^0
\end{aligned}$$

We can continue like this up to any k -th iteration.

$$a = r^0 \vee [t' \times r^1] \vee [(t')^2 \times r^2] \vee \cdots \vee [(t')^k \times r^k] \vee [(t')^{k+1} \times a^{k+1}]$$

or, if we let $(t')^0$ denote the identity matrix e , we have

$$a = \bigvee_{i=1}^k [(t')^i \times r^i] \vee [(t')^{k+1} \times a^{k+1}]$$

We now state a result which will be useful in describing the division algorithm in the image algebra.

Lemma 6.3. *Let $a, b \in F^n$ (be finite vectors). Then we may express the difference of vectors a and b , $a - b$, using the following matrix transform. Define $s \in M_{nn}(-\infty)$ by $s = \text{diag}((b_1)^*, \dots, (b_n)^*) = \text{diag}(-b_1, \dots, -b_n)$ with $-b_i$ denoting the real arithmetic additive inverse of the real number b_i . Then*

$$s \times a = c \in F^n, \text{ where}$$

$$c_i = a_i - b_i, \quad i = 1, \dots, n.$$

Proof:

$$(s \times a)_i = \bigvee_{k=1}^n (s_{ik} + a_k) = s_{ii} = a_i = -b_i + a_i$$

for $i = 1, \dots, n$.

Q.E.D.

We remark that the vector r as defined in Theorem 6.2 can be obtained in the following way. Fix $a \in F^n$, finite. Define $f_a : F^n \rightarrow F^n$ by

$$f_a(x) = y \text{ where } y_i = \begin{cases} 0 & \text{if } a_i > x_i \\ -\infty & \text{otherwise} \end{cases}$$

$$\text{Then for } x = t' \times (\hat{t} \times' a), f_a(x) = \begin{cases} 0 & \text{if } a_i > [t' \times (\hat{t} \times' a)]_i \\ -\infty & \text{if } a_i \leq [t' \times (\hat{t} \times' a)]_i \end{cases}. \text{ However, it is}$$

easily shown that f_a is not an s -lattice homomorphism. For example, choose $n = 2$, a , d , and e as below:

$$a = \begin{bmatrix} 3 \\ -1 \end{bmatrix}, \quad d = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \quad e = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

$$\text{Then } f_a(d \vee e) = f_a\left(\begin{bmatrix} 3 \\ -2 \end{bmatrix}\right) = \begin{bmatrix} -\infty \\ 0 \end{bmatrix}, \text{ but } f_a(d) \vee f_a(e) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \vee \begin{bmatrix} -\infty \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \neq f_a(d \vee e).$$

Thus, according to Theorem 3.1, f_a cannot be represented as a matrix transform. If, however, we go outside of the structure of the minimax algebra, and use image algebra operations in addition to \vee and \boxtimes (or \odot), we can express this transform in a succinct way, as will be demonstrated in the next section.

A dual division algorithm. The duality of the operations of the matrix algebra enable us to describe a dual division algorithm. We omit the proofs, as they are the dual of the proofs given in the previous section.

Lemma 6.4. *Let $\mathbf{a} \in \mathbb{F}^n$ be finite, and $\mathbf{t} \in M_{nn}(+\infty)$ satisfy $S_{+\infty}(t_i) \neq \emptyset \forall i = 1, \dots, n$. Define $\hat{\mathbf{t}}$ by $\hat{\mathbf{t}} \equiv (\mathbf{t}^*)'$. Then both $\mathbf{t}' \times \mathbf{a}$ and $\hat{\mathbf{t}} \times' (\mathbf{t}' \times \mathbf{a})$ are finite, and*

$$\hat{\mathbf{t}} \times' (\mathbf{t}' \times \mathbf{a}) \geq \mathbf{a}$$

Lemma 6.5. (The Dual Division Algorithm) *Let \mathbf{a}, \mathbf{t} satisfy the hypothesis of Lemma 6.4. Then for $\mathbf{q} = \mathbf{t}' \times \mathbf{a}$, and \mathbf{r} defined by*

$$r_i = \begin{cases} a_i & \text{if } a_i < [\hat{\mathbf{t}} \times' (\mathbf{t}' \times \mathbf{a})]_i \\ +\infty & \text{if } a_i = [\hat{\mathbf{t}} \times' (\mathbf{t}' \times \mathbf{a})]_i \end{cases}$$

we have

$$\mathbf{a} = (\hat{\mathbf{t}} \times' \mathbf{q}) \wedge \mathbf{r}.$$

6.2. An Image Algebra Division Algorithm

Using the isomorphism Ψ , we can express these ideas in the image algebra. Let $\hat{\mathbf{t}} \equiv (\mathbf{t}^*)'$ for $\mathbf{t} \in (\mathbb{F}_{-\infty}^X)^X$.

Lemma 6.6. *Let $\mathbf{a} \in \mathbb{F}^X$, $\mathbf{t} \in (\mathbb{F}_{-\infty}^X)^X$ be such that $S_{-\infty}(t_x) \neq \emptyset \forall x \in X$. Then each of $\mathbf{a} \boxtimes \hat{\mathbf{t}}$ and $(\mathbf{a} \boxtimes \hat{\mathbf{t}}) \boxtimes \mathbf{t}'$ are finite, and $\mathbf{a} \geq (\mathbf{a} \boxtimes \hat{\mathbf{t}}) \boxtimes \mathbf{t}'$.*

This next theorem is the counterpart to Lemma 6.3.

Lemma 6.7. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{F}^X$. Then the image $\mathbf{c} = \mathbf{a} - \mathbf{b}$ may be expressed using a template in the following way. Define $\mathbf{s} \in (\mathbb{F}_{-\infty}^X)^X$ by*

$$s_y(x) = \begin{cases} -b(y) & \text{if } x = y \\ -\infty & \text{otherwise} \end{cases}$$

Then $a \boxminus s = a - b$.

Using the lattice characteristic function, it is sometimes the case that we can stay within the lattice operations of \vee and \boxminus and the image algebra operation of $+$ when needing to express a characteristic function. An example of this follows immediately.

Theorem 6.8. The Division Algorithm. *Let a, t satisfy the hypothesis of Lemma 6.6.*

Then for $q = a \boxminus \hat{t}$ and r defined by

$$r = a + \chi_{>0}^\infty(a)[a - ((a \boxminus \hat{t}) \boxminus t')]$$

we have that

$$a = (q \boxminus t') \vee r.$$

Proof: We need to show that $\Psi^{-1}(r)$ matches with our definition of the matrix r in

Theorem 6.2. Let $b = (a \boxminus \hat{t}) \boxminus t'$. Then using Lemma 6.7, $a - b = a \boxminus s$,

where

$$s_y(x) = \begin{cases} -b(y) & \text{if } x = y \\ -\infty & \text{otherwise} \end{cases}$$

Thus, $a - b \geq 0$ implies that $a \boxminus s \geq 0$. Now,

$$\chi_{>0}^\infty(a \boxminus s) = c \in F^X, \quad \text{where} \quad c(x) = \begin{cases} 0 & \text{if } a(x) > b(x) \\ -\infty & \text{if } a(x) = b(x) \end{cases}$$

Thus, at location $x \in X$, the image $r = a + \chi_{>0}^\infty[a \boxminus s]$ has the gray value

$$r(x) = a(x) + c(x) = \begin{cases} a(x) + 0 & \text{if } a(x) > b(x) \\ a(x) + -\infty & \text{if } a(x) = b(x) \end{cases} = \begin{cases} a(x) & \text{if } a(x) > b(x) \\ -\infty & \text{if } a(x) = b(x) \end{cases}$$

Under Ψ , this remainder image is the same as the vector r in Lemma 6.2.

Q.E.D.

Iterating k times on an image a and a template t satisfying the hypothesis of Lemma 6.6, we obtain

$$a = \bigvee_{i=1}^k [(r^i \boxtimes (t')^i) \vee [(a^{k+1} \boxtimes (t')^{k+1})]$$

where any template t raised to the zero-th power, t^0 , is the identity template, e .

In the boolean case, there exists an integer m such that

$$a_m \boxtimes (t')^m = 0$$

so that the expression for a becomes

$$a = \bigvee_{k=0}^m r_k \boxtimes (t')^k$$

One useful application of this result is in data compression. By encoding the r_i 's in run length code, the image can be represented by fewer bits of data, and reconstructed exactly once t is known.

We have the dual Division Algorithm stated in the image algebra also.

Proposition 6.9. *Let $a \in R^X$, $t \in (R_{+\infty}^X)^X$ be such that $S_{-\infty}(t_x) \neq \emptyset \forall x \in X$. For \hat{t} by $\hat{t} \equiv (t^*)'$, we have that each of $a \boxtimes t'$ and $(a \boxtimes t') \boxtimes \hat{t}$ are finite, and $a \leq (a \boxtimes t') \boxtimes \hat{t}$.*

Proposition 6.10. (The Dual Division Algorithm). *Let a, t satisfy the hypothesis of Proposition 6.9. Then for $q = a \boxtimes t'$ and r defined by*

$$r = a \wedge \chi_{<0}^\infty[a - ((a \boxtimes t') \boxtimes \hat{t})]$$

we have that

$$a = (q \boxtimes \hat{t}) \wedge r.$$

CHAPTER 7 TWO EXAMPLES

7.1. An Operations Research Problem Stated in Image Algebra Notation

This section gives a short description of the *transportation problem* in linear programming [63], and provides a translation of the dual transportation problem into image algebra notation. Thus, this provides an example of the use of the isomorphism Ψ^{-1} .

Let m producers and n consumers of some commodity be given. Let p_i denote the production capacity of producer i , d_j denote the demand of consumer j , and c_{ij} denote the cost of transporting one unit of commodity from producer i to consumer j . The problem is to determine how much commodity to ship from each producer to each consumer so that consumer demands are met, production capacities are not exceeded, and transportation costs are minimized. This can be formulated as a linear programming (LP) problem, which we state as follows.

Let z_{ij} be the number of units of commodity to be shipped from producer i to consumer j . Then the total transportation cost

$$\sum_{i=1}^m \sum_{j=1}^n c_{ij} z_{ij}$$

is to be minimized. To stay within production capacity, we also must have

$$\sum_{j=1}^n z_{ij} \leq p_i, \quad i = 1, \dots, m$$

and to satisfy consumer demands we must have

$$\sum_{i=1}^m z_{ij} \geq d_j, \quad j = 1, \dots, n.$$

Thus the LP problem is to

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^m \sum_{j=1}^n c_{ij} z_{ij} \\ \text{Subject to} \quad & -\sum_{j=1}^n z_{ij} \geq -p_i, \quad i = 1, \dots, m \quad (7-1) \end{aligned}$$

$$\sum_{i=1}^m z_{ij} \geq d_j, \quad j = 1, \dots, n, \text{ and} \quad (7-2)$$

$$z_{ij} \geq 0 \quad \text{for all } i, j.$$

Let x_i be the dual variable associated with the i -th constraint in (7-1), and y_j the dual variable associated with the j -th constraint in (7-2). Then the dual transportation problem is [64]

$$\begin{aligned} \text{Maximize} \quad & -\sum_{i=1}^m p_i x_i + \sum_{j=1}^n d_j y_j \\ \text{Subject to} \quad & -x_i + y_j \leq c_{ij} \quad \text{for all } i, j \\ & x_i \geq 0, y_j \geq 0 \quad \text{for all } i, j. \end{aligned}$$

This is equivalent to solving

$$\begin{aligned} \text{Minimize} \quad & -\left[-\sum_{i=1}^m p_i x_i + \sum_{j=1}^n d_j y_j\right] \\ \text{Subject to} \quad & -x_i + y_j \leq c_{ij} \quad \text{for all } i, j \\ & x_i \geq 0, y_j \geq 0 \quad \text{for all } i, j, \end{aligned}$$

which is

$$\begin{array}{ll}
\text{Minimize} & \sum_{i=1}^m p_i x_i - \sum_{j=1}^n d_j y_j \\
\text{Subject to} & -x_i + y_j \leq c_{ij} \quad \text{for all } i, j \\
& x_i \geq 0, y_j \geq 0 \quad \text{for all } i, j.
\end{array}$$

Make a change of variables by letting

$$v_j = -y_j \quad \text{and} \quad u_i = -x_i, \quad \text{for all } i, j.$$

Then we have the equivalent dual LP problem

$$\begin{array}{ll}
\text{Minimize} & \sum_{j=1}^n d_j v_j - \sum_{i=1}^m p_i u_i \\
\text{Subject to} & u_i - v_j \leq c_{ij} \quad \text{for all } i, j \\
& u_i \leq 0, v_j \leq 0 \quad \text{for all } i, j.
\end{array}$$

Using the theory of *complementary slacks* [64], if we assume that the producers p_i have value $p_i > 0$ for all i , then we can be guaranteed that for each $i = 1, \dots, m$, there exists at least one $j \in \{1, \dots, n\}$ such that

$$u_i - v_j = c_{ij},$$

and, hence,

$$u_i = \min_{j=1}^n \{c_{ij} + v_j\}, \quad (7-3)$$

where $\mathbf{u} = (u_1, \dots, u_m)$ and $\mathbf{v} = (v_1, \dots, v_n)$ are optimal feasible solutions. We can rewrite (7-3) in vector notation, as

$$\mathbf{u} = \mathbf{C} \times' \mathbf{v},$$

and $u_i, v_j \leq 0$.

To formulate this problem in context of the image algebra, we define \mathbf{X} and \mathbf{Y} to be non-empty, finite coordinate sets, $|\mathbf{X}| = m$, $|\mathbf{Y}| = n$. Define $\mathbf{d} \in (\mathbf{R}_{\pm\infty}^{\mathbf{Y}})^{\mathbf{Y}}$ by

$$d_{y_i}(y_j) = \begin{cases} d_i & i = j \\ -\infty & \text{otherwise} \end{cases},$$

and define $p \in (R_{\pm\infty}^X)^X$ by

$$p_{x_i}(x_j) = \begin{cases} p_i & i = j \\ -\infty & \text{otherwise} \end{cases}.$$

Define $a \in R_{\pm\infty}^X$, $b \in R_{\pm\infty}^Y$ by

$$a(x_i) = u_i \text{ (the variable)}$$

$$b(x_i) = v_i \text{ (the variable)}.$$

Then we have:

LP	Image Algebra
$\sum_{j=1}^n d_j v_j$	$\sum (b \boxtimes d)$
$\sum_{i=1}^m p_i u_i$	$\sum (a \boxtimes p)$

Now, define $t \in (R_{\pm\infty}^Y)^X$ by

$$t_{x_i}(y_j) = c_{ij}$$

We have the equation $u = C \times' v$ translates as $a = b \boxtimes t$. Thus, in image algebra notation, the dual LP problem is

$$\text{Minimize} \quad \sum (b \boxtimes d) - \sum (a \boxtimes p)$$

$$\text{Subject to} \quad a = b \boxtimes t$$

$$a \leq 0, b \leq 0$$

7.2. An Image Complexity Measure

This section presents an *image complexity measure*, a term used in image processing to describe any method which provides a quantitative measure of some feature or set of features in an image. Image complexity measures are used either as a pre-processing step in which the measures help direct the selection of the next processing step, or in conjunction with other information derived from the image to identify objects of interest.

The measure investigated [65] is based on a method discussed by Mandelbrot [66] for curve length measurements. The original algorithm was modified and translated into image algebra. The measure itself consists of a graph which in theory gives an indication of the rate of change of variation in the gray level surface. The algorithm for computing the measure is presented, followed by a discussion of an application to 12 outdoor images.

The general approach of the algorithm is to make successive approximations of the area of a gray level surface, and then plot the approximations using a log-log scale. The log-log scale is purported to allow a better visual inspection of the information contained in the graph.

Consider all points with distance to the gray level surface of no more than k . These points form a blanket of thickness $2k$, and the suggested surface area $A(k)$ of the gray level surface is the volume of the blanket divided by $2k$. Here we have $A(k)$ increasing as k decreases.

To begin the computation of the surface area for $k = 1, 2, \dots$, an upper surface u_k and a lower surface b_k are defined iteratively in the following manner. Let a be the input image. Let

$$u_0 = a, \quad b_0 = a$$

Then define u_k and b_k for $k = 1, 2, \dots$, by

$$u_k = u_{k-1} \boxtimes t$$

$$b_k = b_{k-1} \boxtimes -t$$

where

$$t = \begin{array}{|c|c|c|} \hline & 0 & \\ \hline 0 & 1 & 0 \\ \hline & 0 & \\ \hline \end{array}$$

The volume $v(k)$ of the "blanket" between the upper and lower surfaces is calculated for each k by computing

$$p_1(k) = u_k \oplus s, \quad q_1(k) = b_k \oplus (-s)$$

where

$$s = \begin{array}{|c|c|} \hline 0.67 & 0.33 \\ \hline 0.33 & 0.67 \\ \hline \end{array}$$

Let $v_1(k) = \sum [p_1(k) + q_1(k)]$.

This method of estimating the volume was derived using elementary calculus. We explain the method for calculating the volume between the upper surface and the coordinate set \mathbf{X} . The volume between the lower surface and \mathbf{X} is found in a similar way. Given four pixel locations in \mathbf{X} , (i,j) , $(i,j+1)$, $(i+1,j)$, and $(i+1,j+1)$, a box was constructed from the eight points in \mathbf{R}^3 corresponding to the four gray values $u_k(i,j)$, $u_k(i,j+1)$, $u_k(i+1,j)$, $u_k(i+1,j+1)$, and the four given pixels. Drawing a line from $u_k(i,j)$ to $u_k(i+1,j+1)$ and a line from (i,j) to $(i+1,j+1)$, the volume of the triangular column determined by the six points $u_k(i,j)$, $u_k(i+1,j)$, $u_k(i+1,j+1)$, (i,j) , $(i+1,j)$, and $(i+1,j+1)$ was found using elementary methods from calculus. Similarly, the volume of the triangular column determined by the six points $u_k(i,j)$, $u_k(i,j+1)$, $u_k(i+1,j+1)$, (i,j) , $(i,j+1)$, and $(i+1,j+1)$ was determined. The volume

of the two pieces are added together to give an estimate to the volume of the box determined by the eight initial points. This is done over the entire coordinate set \mathbf{X} , and all volumes added together to give an estimate of the volume between \mathbf{X} and the gray value surface u_k . The method was expressed using the image algebra operation \oplus and an invariant template, omitting the boundary effects. Using this approach, the volume is overestimated, so it is corrected by applying a variant template \mathbf{w} effective only on the edge pixels. Define \mathbf{w} by

$$\mathbf{w}_x = \begin{array}{|c|c|} \hline \text{0.33} & \text{0.67} \\ \hline \end{array}$$

if x is a top edge pixel and not the top right corner pixel,

$$\mathbf{w}_x = \begin{array}{|c|} \hline \text{0.33} \\ \hline \end{array}$$

if x is the top right corner pixel,

$$\mathbf{w}_x = \begin{array}{|c|} \hline \text{0.67} \\ \hline \text{0.33} \\ \hline \end{array}$$

if x is a right edge pixel but not the top right corner pixel, and $\mathbf{w}_x = \mathbf{0}$, if x is otherwise.

To correct for the extra volume added in on the edge pixels, we calculate

$$\mathbf{p}_2(k) = \mathbf{u}_k \oplus (-\mathbf{w}), \quad \mathbf{q}_2(k) = \mathbf{b}_k \oplus \mathbf{w}$$

and let $\text{volerr}(k) = \sum [\mathbf{p}_2(k) + \mathbf{q}_2(k)]$. The correct volume $v(k)$ is

$$v(k) = v_1(k) + \text{volerr}(k).$$

The approximated surface area is

$$\text{area}(k) = \frac{v(k)}{2k}.$$

The rate of change of $\log(\text{area}(k))$ with respect to $\log(k)$ contains important information about the image. The slope $S(k)$ of $\text{area}(k)$ versus k is computed on a log-log scale for each k

by finding the best fitting straight line through the three points

$$(\log(k-1), \log(\text{area}(k-1))), (\log(k), \log(\text{area}(k))), (\log(k+1), \log(\text{area}(k+1))).$$

The graph of $S(k)$ versus k is called the *signature* of the image. We can also calculate a signature for the case where the array \mathbf{X} represents the bottom surface and \mathbf{u}_k the upper surface. We call this the *upper signature*. Similarly, the signature which is calculated using $\{\mathbf{b}_k\}$ for the lower surfaces and \mathbf{X} for the upper surfaces is called the *lower signature*.

This algorithm was run on 12 images. For each image, we calculated the input image, \mathbf{u}_i , \mathbf{b}_i , $i = 1, \dots, 50$, and the graph of the upper and lower signatures.

As k increases, regions of pixels initially having the highest gray values decrease in size in the images \mathbf{b}_k . However, as k increases, the images \mathbf{u}_k shrink regions having lower gray values. In theory, this asymmetry can be taken advantage of. Roughly, the lower signature represents the shape of objects with high gray values, and the upper signature represents the distribution of objects throughout the image. The images to which we applied this method were infrared, so we were mainly interested in the lower signatures.

The magnitude of the curve $S(k)$ is related to the information lost on objects with details less than k in size. The more gray level variation at distance k , the higher the values for $S(k)$. Thus, if at small k $S(k)$ is large, then there is "high-frequency" gray level variations, and if at large k $S(k)$ is large, then we have "low-frequency" gray level variations. The curve $S(k)$ thus gives us information about the rate of change of variations in the gray level surface.

After running the program on a dozen images, we have concluded that this algorithm is too sensitive to the great variance in outdoor scenery. For example, an image which has a background of trees and no targets, and an image which has two distinct targets and no trees as background have similar graphs for the lower signatures. While the lower signature

represents more of the shape of the hot objects (areas with high gray values) in the image, in one image we have no hot objects while in the other, there are two distinct hot objects. As another example, in two other images we have a target with a road and a field as background, yet the graphs for the upper signatures for these images have a very distinct difference. The theory suggests that upper signatures should represent similar targets, but we cannot draw that conclusion from this data. The examples given in the original paper [64] are of a very regular texture and are presented in a controlled environment. It is very likely that the controlled environment in which the data was taken is one reason why this algorithm was successful for those authors.

The initial goal of investigating this type of complexity measure was that these graphs would hopefully give a measure of gray level variation within an image and help in choosing a more effective edge operator. If an image has a high incidence of gray level variation at small values of k , then it is reasonable to assume a more sensitive mask, such as the gradient mask, would give better results. Otherwise, if an image had small values of $S(k)$ at small values of k , then computation time could be saved by using a Sobel operator instead of a computationally intensive edge operator such as the Kirsch. Unfortunately, the algorithm did not produce data that leads to this conclusion.

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

We have shown that the establishment of an isomorphism between a lattice-based matrix theory and the image algebra provides a powerful tool in image processing which was not available previous to this research. Just as linear transforms are able to be represented as matrices within the structure of linear algebra, lattice transforms are able to be represented as matrices within the structure of the minimax algebra, and thus all mathematical results of the minimax algebra are applicable to solving problems in image processing.

In particular, we have shown that

- (1) Many notions encountered in investigating linear transforms, such as the eigenvalue/eigenvector problem, rank, linear independence, and solutions to systems of equations, have their counterpart in image algebra. The use of these notions to solving specific image processing problems remains to be seen.
- (2) The image algebra is useful as a model in mapping a class of non-linear transforms to parallel architectures. It is feasible to map any arbitrary lattice transform to most parallel architectures; that is, given a network of processors that are interconnected by communication links, a lattice transform has a weak decomposition into a product of lattice transforms that are each local with respect to the network in and only if every pair of processors has a two-way path of communication between them. There is at most one weak (pointwise) operation of maximum, with the remaining operations being convolutions. The pointwise maximum which represents the fact that a small amount of storage is needed to compute the transform is not an unreasonable restriction.

- (3) Mathematical morphology is a special subclass of the lattice transforms, namely the transforms corresponding to invariant templates from \mathbf{X} to \mathbf{X} having the target pixel as part of their support at each pixel. Thus, the lattice subalgebra of the image algebra generalizes mathematical morphology. The image algebra is a clear and simple mathematical representation of morphological concepts, avoiding the cumbersome notion of *umbra*, for example.
- (4) The minimax algebra lends itself well to the investigation of matrix decomposition techniques. A weak decomposition was shown for any matrix lattice transform. Thus, the application of a matrix product for any purpose can be implemented in parallel. This includes problems from operations research.
- (5) Although the minimax algebra is not a Euclidean domain, the representation of a type of division algorithm was found. In boolean images, there are clear applications of these types of skeletonizing techniques to image compression.

We now list some suggestions for further research. The author continues to pursue this general area of research, and would appreciate being informed of any results obtained related to the following problems.

Unlike linear algebra, the minimax algebra is relatively unknown. The matrix properties developed so far in the literature [39] were meant primarily for use in operations research problems. A wealth of mathematical results, with possible applications to real world problems, are obviously obtainable from investigating other uses of the minimax algebra. The matrix decomposition techniques presented in this dissertation are a clear example. Specific problems are: investigate Chapter 3's statements for applications to solving image processing problems. What specific applications does the eigenproblem have in image processing? The statements in Chapter 3 are only a few of the minimax algebra properties that were mapped

to image algebra. Investigate other properties in the *Minimax Algebra* book which may have uses in image processing. Is there a minimax transform equivalent to the linear Fourier transform? What are its uses? What other types of linear transforms have a minimax equivalent? Can other techniques in linear algebra, such as the SVD, Kronecker products, etc. be described in a similar way in the minimax algebra? Develop techniques for decomposition of square matrices that are block toeplitz with toeplitz blocks, with non-null diagonal entries. As mentioned in Chapter 4, this type of matrix corresponds to the concept of a structuring element in mathematical morphology, and to a class of invariant templates in the image algebra.

Another question concerns the existence theorem, Theorem 5.24. Does there exist a local decomposition that is not weak, that is, a decomposition with no operation of \vee ? This would relieve the storage requirement associated with \vee . Also, is there a similar existence theorem with the digraph $D(W)$ replacing the graph $G(W)$? Although the existence theorem given in Chapter 5 is constructive, one would rarely implement the decomposition in this manner. Describe a general decomposition technique giving local templates and which is minimal with respect to some criterion, such as the number of factors in the decomposition. Also, what type of applications to operations research does the decomposition theorem have?

The continuous cases for the image algebra operands \oplus , \boxtimes , and \odot have yet to be well established. Establish the mathematical foundations for the continuous counterparts involving the lattice operations \boxtimes , \odot , and \vee .

In his dissertation Gader asks the question about relating circulant templates under \boxtimes or \odot to group algebras. Will the formal minimax matrix relationship established with the image algebra make the pursuit of this particular question any easier?

Is there a practical use for the division algorithm? Generalize the division algorithm to a sequence of templates t_i replacing the single t .

REFERENCES

1. J. von Neumann, "The General Logical Theory of Automata," in *Cerebral Mechanism in Behavior: The Hixon Symposium*, Wiley and Sons, New York (1951).
2. S.H. Unger, "A Computer Oriented Toward Spatial Problems," *Proc. IRE* **46** (1958), 1744-1750.
3. K.E. Batcher, "Design of a Massively Parallel Processor," *IEEE Trans. Computers* **29**(9) (1980), 836-840.
4. M.J.B. Duff, "CLIP4," in *Special Computer Architectures for Pattern Processing*, ed. K.S. Fu, CRC Press, Boca Raton, FL (1982).
5. T.J. Fountain, K.N. Matthews, and M.J.B. Duff, "The CLIP7A Image Processor," *IEEE Pattern Analysis and Machine Intelligence* **10**(3) (May 1988).
6. L. Uhr, "Pyramid Multi-Computer Structures, and Augmented Pyramids," in *Computing Structures for Image Processing*, ed. M.J.B. Duff, Academic Press, London (1983).
7. W.D. Hillis, *The Connection Machine*, The MIT Press, Cambridge, MA (1985).
8. J.C. Klein and J. Serra, "The Texture Analyzer," *J. Micros* **95** (1972), 349-356.
9. S.R. Sternberg, "Biomedical Image Processing," *Computer* **16**(1) (January 1983), 22-34.
10. D.L. McCubbrey and R.M. Loughheed, "Morphological Image Analysis Using a Raster Pipeline Processor," pp. 444-452 in *IEEE Comp. Soc. Workshop on Comp. Arch. for Patt. Analysis and Image Database Management*, Miami Beach, FL (1985).
11. E. Cloud and W. Holsztynski, "Higher Efficiency for Parallel Processors," pp. 416-422 in *Proceedings IEEE Southcon 84*, Orlando, FL (March, 1984).
12. H. Minkowski, *Gesammelte Abhandlungen*, Teubner Verlag, Leipzig-Berlin (1911).
13. H. Minkowski, "Volumen und Oberflache," *Mathematische Annalen* **57** (1903), 447-495.
14. Hadwiger, *Vorlesungen uber Inhalt, Oberflache und Isoperimetrie*, Springer-Verlag, Berlin (1957).
15. G. Matheron, *Random Sets and Integral Geometry*, Wiley, New York (1975).
16. J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, London (1982).
17. T.R. Crimmins and W.M. Brown, "Image Algebra and Automatic Shape Recognition," *IEEE Trans. Aerospace and Elec. Systems* **AES-21**(1) (Jan. 1985), 60-69.
18. R.M. Haralick, L. Shapiro, and J. Lee, "Morphological Edge Detection," *IEEE Journal of Robotics and Automation* **RA-3**(1) (April 1987), 142-157.
19. R.M. Haralick, S.R. Sternberg, and X. Zhuang, "Image Analysis Using Mathematical Morphology: Part I," *IEEE Trans. on Patt. Analysis and Mach. Intelligence* **PAMI-9**(4) (July 1987), 532-550.

20. P. Maragos and R.W. Schafer, "Morphological Skeleton Representation and Coding of Binary Images," *IEEE Trans. Acoustics, Speech, and Signal Proc.* ASSP-34(5) (Oct. 1986), 1228-1244.
21. P. Maragos and R.W. Schafer, "Morphological Filters Part I: Their Set-Theoretic Analysis and Relations to Linear Shift-Invariant Filters," *IEEE Trans. Acoustics, Speech, and Signal Proc.* ASSP-35 (Aug. 1987), 1153-1169.
22. P. Maragos and R.W. Schafer, "Morphological Filters Part II: Their Relations to Median, Order-Statistic, and Stack Filters," *IEEE Trans. Acoustics, Speech, and Signal Proc.* ASSP-35 (Aug. 1987), 1170-1184.
23. S.R. Sternberg, "Language and Architecture for Parallel Image Processing," in *Proc. of the Conf. on Pattern Rec. in Practice*, Amsterdam (May 1980).
24. S.R. Sternberg, "Overview of Image Algebra and Related Issues," in *Integrated Technology for Parallel Image Processing*, ed. S. Levialdi, Academic Press, London (1985).
25. P. Maragos, "A Unified Theory of Translation-Invariant Systems With Applications to Morphological Analysis and Coding of Images," Ph.D. thesis, Georgia Inst. Tech., Atlanta (1985).
26. P.E. Miller, "Development of a Mathematical Structure for Image Processing," Optical Division Tech. Report, Perkin-Elmer (1983).
27. R. Hockney and C. Jesshope, *Parallel Computers : Architecture, Programming, and Algorithms*, Adam Hilger, Bristol (1981).
28. J.T. Schwartz, "A Taxonomic Table of Parallel Computers, Based on 55 Designs," Technical Report, Courant Institute of Mathematical Sciences, New York University (1983).
29. A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, NJ (1975).
30. S.R. Sternberg, "Languages and Architectures for Parallel Image Processing," in *Pattern Recognition in Practice*, North-Holland Publishing Co., New York, NY (1980).
31. O. Ersoy, "Semisystolic Array Implementation of Circular, Skew Circular, and Linear Convolutions," *IEEE Trans. on Computers* C-34(2) (February 1985), 190-196.
32. G.X. Ritter, J.N. Wilson, J.L. Davidson, "Image Algebra: An Overview," to appear in *Computer Vision, Graphics, and Image Processing*, (1989).
33. P.D. Gader, "Image Algebra Techniques for Parallel Computation of Discrete Fourier Transforms and General Linear Transforms," Ph.D. Dissertation, University of Florida, Gainesville, FL (1986).
34. G.X. Ritter, J.N. Wilson, and J.L. Davidson, "Standard Image Processing Algebra Document Phase II," TR (7) Image Algebra Project, F08635-84-C-0295, Eglin AFB, FL (1987).
35. J. Serra, *Image Analysis and Mathematical Morphology, Volume 2: Theoretical Advances*, Academic Press, New York (1988).
36. Birkhoff, *Lattice Theory*, American Mathematical Society, Providence, RI (1984).
37. R. Cuninghame-Green, "Process Synchronisation in Steelworks - a Problem of Feasibility," pp. 323-328 in *Proc. 2nd Int. Conf. on Oper. Research*, ed. Banbury, English University Press, London (1960).

38. R. Cuninghame-Green, "Describing Industrial Processes with Interference and Approximating their Steady-State Behaviour," *Oper. Research Quart.* 13 (1962), 95-100.
39. R. Cuninghame-Green, *Minimax Algebra: Lecture Notes in Economics and Mathematical Systems 166*, Springer-Verlag, New York (1979).
40. J.E. Cohen, "Subadditivity, Generalized Products of Random Matrices and Operations Research," *SIAM Review*, (March 1988), 69-86.
41. G. Birkhoff and J. Lipson, "Heterogeneous Algebras," *J. Combinatorial Theory* 8 (1970), 115-133.
42. G.X. Ritter, M.A. Shrader-Frechette, and J.N. Wilson, "Image Algebra: A Rigorous and Translucent Way of Expressing All Image Processing Operations," in *Proc. of the 1987 SPIE Tech. Symp. Southeast on Optics, Elec.-Opt., and Sensors*, Orlando, FL (May 1987).
43. G.X. Ritter and J.N. Wilson, "Image Algebra: A Unified Approach to Image Processing," in *Proceedings of the SPIE Medical Imaging Conference*, Newport Beach, CA (February 1987).
44. P.D. Gader, "Necessary and Sufficient Conditions for the Existence of Local Matrix Decompositions," *SIAM Journal on Matrix Analysis and Applications*, (July 1988), 305-313.
45. G.X. Ritter and P.D. Gader, "Image Algebra Techniques for Parallel Image Processing," *Journal of Parallel and Distributed Computing* 4(5) (March 1987), 7-44.
46. A. Shimbel, "Structure in Communication Nets," pp. 119-203 in *Proc. Symp. on Information Networks*, Polytechnic Institute of Brooklyn (1954).
47. B. Giffler, "Mathematical Solution of Production Planning and Scheduling Problems," IBM ASD Tech. Rep. (1960).
48. V. Peteanu, "An Algebra of the Optimal Path in Networks," *Mathematica* 9 (1967), 335-342.
49. C. Benzaken, "Structures Algebra des Cheminements," pp. 40-57 in *Network and Switching Theory*, ed. Biorci, Academic Pres (1968).
50. B Carre, "An Algebra for Network Routing Problems," *J. Inst. Math. Appl.* 7 (1971), 273-294.
51. R.C. Backhouse and B Carre, "Regular Algebra Applied to Path-Finding Problems," *J. Inst. Math. Appl.* 15 (1975), 161-186.
52. G.X. Ritter, M.A. Shrader-Frechette, and P.D. Gader, "Image Algebra Tutorial, Version I," TR Image Algebra Project, F08635-84-C-0295, Eglin AFB, FL (1985).
53. M. Yoeli, "A Note on a Generalization of Boolean Matrix Theory," *Am. Math. Monthly* 68 (1961), 552-557.
54. G. Matheron, *Elements pour une theorie des milieux poreux* 1967.
55. J. Serra, "Introduction a la Morphologie Mathematique," Booklet No. 3, Cahiers du Centre de Morphologie Mathematique, Fontainebleau, France (1969).
56. S. Sternberg, "Cellular Computers and Biomedical Image Processing," in *Proc. U.S. - France Seminar on Biomed. Image Processing*, Grenoble, France (1980).

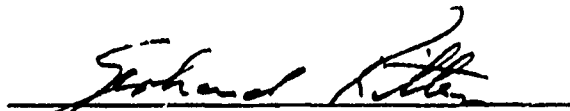
57. J. Serra, "Morphologie pour les fonctions "a peu pres en tout ou rien"," Technical Report, Cahiers du Centre de Morphologie Mathematique, Fontainebleau, France (1975).
58. F. Meyer, "Iterative image transformation for an automatic screening of cervical smears," *Journal of Histochem. and Cytochem.* 27(1) (1978), 128-135.
59. S.R. Sternberg, "Esoteric Iterative Algorithms," in *II Internat'l. Conf. on Image Analysis and Processing*, Brindisi, Italy (Nov. 1982).
60. G.X. Ritter, J.N. Wilson, and J.L. Davidson, "Image Algebra Application to Image Measurement and Feature Extraction," in *Proc. of the 1989 SPIE OE/LASE 1989 Optics, Elec.-Optics, and Laser Appl. in Sci. and Eng.*, Los Angeles (Jan. 1989).
61. G.X. Ritter, Dong Li, "Template Decomposition and Image Algebra," UF-CIS Technical Report, Dept. of Comp. and Info. Sci., Univ. of Florida, Gainesville, FL (1989).
62. J.B. Fraleigh, *A First Course in Abstract Algebra*, Addison-Wesley, Reading, MA (1968).
63. P.E. Miller, "An Investigation of Boolean Image Neighborhood Transformations," Ph.D. dissertation, Ohio State University (1978).
64. K.G. Murty, *Linear and Combinatorial Programming*, John Wiley, New York (1976).
65. S. Peleg and et. al., "Multiple Resolution Texture Analysis and Classification," TR, Center for Automation Research, University of Maryland, College Park, MD (July 1983).
66. B.B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco (1983).

BIOGRAPHICAL SKETCH

Jennifer L. Davidson was born in Dayton, Ohio, and grew up in the Washington, D.C. area. She received the B.A. in physics from Mount Holyoke College, Massachusetts, and the M.S. in mathematics from the University of Florida in 1986. Her research interests include applied mathematics, image processing, and computer vision, and she is the author of over 17 research papers including a journal publication.

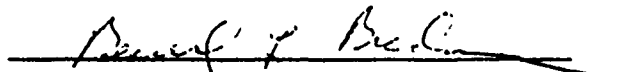
Ms. Davidson is a member of the IEEE Computer Society, Society for Photo-Optical Instrumentation Engineers, Society of Industrial and Applied Mathematics, Mathematical Association of America, and the American Mathematical Society.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



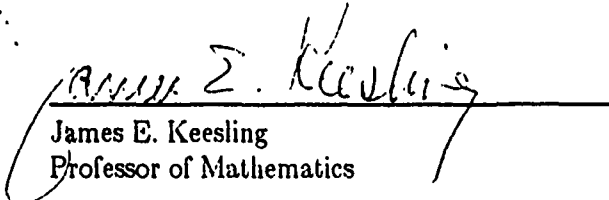
Gerhard X. Ritter, Chairman
Professor of Mathematics
Professor of Computer and Information Science

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



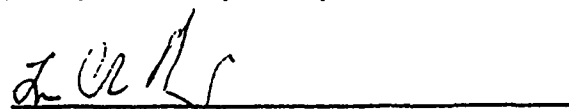
Beverly L. Breckner
Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



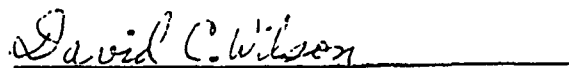
James E. Keesling
Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Li-Chien Shen
Associate Professor of Mathematics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



David C. Wilson
Professor of Mathematics

APPENDIX F

DOD FELLOWSHIP FORMS



Universal Energy Systems, Inc.

---Date-----

UES Project 789
DOD Fellows

ADDRESS

SUBJECT: Contract F49620-86-C-0127/SB5861-0436
Proposed Subcontract S-789-DOD-000

Dear _____:

Enclosed are an original and one copy of the subject subcontract agreement. The total price is \$ 17,500.

This Subcontract Agreement must be signed by an official of your organization who is authorized to bind the University of _____ in a legal contract.

Please sign and return the original agreement as soon as possible, but no later than 10 days after the date of this letter. This agreement will be signed, and a copy of the fully executed document will be returned to you for your files.

If you have any contractual questions, please call me. For technical questions, please contact Mr. Rodney C. Darrah, Project Manager.

Sincerely,

UNIVERSAL ENERGY SYSTEMS, INC.

Susan K. Espy
Contract Administrator

xc: DCASMA/Dayton (M. Hughes)
UES Project Manager
UES Suspense (Ltr Only)
Fellow

SKE/mt
2329s

SUBCONTRACT AGREEMENT		Page 1 of 7
Subcontract No.: S-789-DOD-000	Prime Contract No.: F49620-86-C-0127 SB5861-0436	Certified For National Defense Use Under DPAS (15 CFR 350). PRIORITY RATING: DO C-7
Issued by: Universal Energy Systems, Inc. 4401 Dayton-Xenia Road Dayton, OH 45432 Project Manager: Rod Darrah		Subcontractor:
Contract Type: FIXED PRICE	Total Amount: \$17,500.00 OPTION I - 22,000 OPTION II - 23,000	
Subcontract Effective Date: 1 September 1989	Security Classification: U Date of DD Form 254: N/A	

TABLE OF CONTENTS

(X)	SEC	ITEM	PAGE
		PART I-SCHEDULE	
X	A	Contract Form	2
X	B	Supplies/Services & Prices	2-3
X	C	Description/Specifications	3
X	D	Packaging & Marking	3
X	E	Inspection & Acceptance	3
X	F	Deliveries or Performance	3-4
X	G	Administration Data	4
X	H	Special Provisions	5-7
		PART II - GENERAL PROVISIONS	
X	I	General Provisions	7-8
		Part III - LIST OF ATTACHMENTS	
X	J	Documents, Exhibits, Attachments	8

Subcontractor agrees to furnish and deliver all items or perform all the services set forth or otherwise identified in this subcontract agreement for the consideration stated.

SUBCONTRACTOR: _____

By: _____ Date _____
 (Signature of Person Authorized to Sign) Signed: _____

Name & Title: _____

UNIVERSAL ENERGY SYSTEMS, INC.:

By: _____ Date _____
 (Signature) Signed: _____

Name & Title: Rodney C. Darrah, Vice President, Scientific Services

SECTION A
CONTRACT FORM

1. This Subcontract Agreement is entered into as of the 1st day of September 1989, between Universal Energy Systems, Inc., an Ohio corporation with offices in Dayton, Ohio, (hereinafter referred to as UES) and the (University) (hereinafter referred to as Subcontractor).

Whereas UES has entered into prime contract No. F49620-86-C-0127 (SB5861-0436) with the United States Government pursuant to which UES is obligated to furnish Services and Related Personnel to conduct special studies involving U.S. Graduate Students pursuing Doctoral Degrees in Air Force Research Activities, and

Whereas UES desires to have Subcontractor perform a portion of the work and Subcontractor desires to assume the obligation to perform such portion of the work, subject to the terms, conditions, and provisions of this subcontract agreement:

Now therefore, in consideration of the foregoing and the covenants herein contained, the parties agree as set forth herein.

2. Work Location. Work will be performed at the (University) in support of the Department of Defense National Defense Science & Engineering Graduate Fellowship Program and (Laboratory).

SECTION B
SUPPLIES, SERVICES, AND PRICES

1. The following services shall be provided:

Educational and research facilities, academic guidance and coursework in support of a Special Study in (Research). (Fellow) is assigned to the (Laboratory) at (Air Force Base). (Mentor) is the Laboratory Advisor. (Advisor), Department of _____, is the Academic/Dissertation Advisor.

2. For services rendered, as herein requested, UES will pay compensation as follows:

Fellow's Stipends:	\$14,000 (\$10,500 Academic Year-\$3,500 Summer Session)
In lieu of Fellow's Tuition and Fees:	\$ 6,000 (First Year)(12 months)
Fellowship Processing Fee:	\$ <u>1,000</u> (First Year)(12 months)
TOTAL	\$21,000

- 3a. Option I - For second year services rendered as herein requested, UES will pay compensation as follows:

Fellow's Stipends:	\$15,000 (\$11,250 Academic Year-\$3,750 Summer Session)
In lieu of Fellow's Tuition and Fees:	\$ 6,000 (Second Year)(12 months)
Fellowship Processing Fee:	\$ 1,000 (Second Year)(12 months)
TOTAL	\$22,000

- b. Option II - For third year services rendered as herein requested, UES shall pay compensation as follows:

Fellow's Stipends:	\$16,000 (\$12,000 Academic Year-\$4,000 Summer Session)
In lieu of Fellow's Tuition and Fees:	\$ 6,000 (Third Year)(12 months)
Fellowship Processing Fee:	\$ 1,000 (Third Year)(12 months)
TOTAL	\$23,000

SECTION C
DESCRIPTION/SPECIFICATIONS

1. The subcontractor as an independent subcontractor, and not as an agent of the contractor, shall perform research or services as set forth in the Fellow's letter, "National Defense Science and Engineering Graduate Fellowship Conditions Form, Additional Opportunities Under the Air Force Laboratory Sponsorship and the Air Force Laboratory Sponsorship Response Form," dated _____, Attachment 1 hereto.

SECTION D
PACKAGING & MARKING

N/A

SECTION E
INSPECTION & ACCEPTANCE

N/A

SECTION F
DELIVERABLES AND PERFORMANCE PERIOD

1. Period of performance is from 1 September 1989 through 31 August 1990. This Subcontract Agreement shall terminate at 12:01 a.m. on 1 September 1990 unless extended by written agreement. The first year of the fellowship tenure is to be completed on or before 31 August 1990.

2. Deliverables are as follows:

- a. Statement of Progress: To be submitted by the Fellow's Academic Advisor through the Subcontractor at the completion of each academic term. (Forms at Attachment 2 hereto).
- b. Publications Prepared By the Fellow: Two (2) copies of any publications prepared by the Fellow to be submitted by the Fellow through the subcontractor.
- c. Dissertation: Complete copy of dissertation covering area of (Research) to be submitted by the Fellow through the Subcontractor.
- d. Transcripts: To be submitted by the Fellow through the Subcontractor at the completion of each academic term.

3. All deliverables shall be delivered to the UES office located at 4401 Dayton-Xenia Road, Dayton, Ohio, 45432, ATTN: Mr. Rodney C. Darrah.

SECTION G
ADMINISTRATIVE DATA

1. PAYMENT:

- a. Initial payment for the Academic Year will be made by UES to the Subcontractor within 45 days after the date of UES's signature on page 1 of this Subcontract. This initial sum, specified in Section H, paragraph 9, includes payment in lieu of tuition for the Academic Year and summer term 1989-1990 (12 months), stipend for Academic Year and summer term 1989-1990 (12 months), and \$1,000.00 for the Fellowship Processing Fee.
- b. Payment for each subsequent summer term and Academic years will be made by UES to the Subcontractor within 45 days after execution of the modification to fund. Payments will include stipend for the period.

SECTION H
SPECIAL PROVISIONS

1. In the event Subcontractor should violate the terms of this subcontract agreement UES, at its option, may terminate this subcontract agreement upon verbal notification to the subcontractor which will be confirmed in writing. UES shall be under no obligation except to pay Subcontractor such compensation as Subcontractor may be entitled to receive up to the time of such termination.
2. The Fellow is expected to complete all doctoral requirements within three (3) years from 1 September 1989.
3. As set forth in Section B.2 above payment of \$6,000 is made in lieu of yearly (12 month) tuition and fees.
4. The normal tenure of this fellowship is twelve (12) months each fellowship year. The twelve (12) month tenure may be reduced to no less than nine (9) months with forfeiture of stipend for the remaining three (3) months of the fellowship year. In the event this occurs this subcontract will be modified to decrease the stipend set forth in Section B.2 above.
5. The availability of funding for years two and three is contingent on a) certification to UES by the (University) that satisfactory academic progress toward a Ph.D in (Research) is being made by (Fellow), and b) the availability of appropriated Air Force funds for continued support. If this Fellowship is continued beyond the first year this Subcontract will be amended accordingly. If amended, the second year stipend will be \$15,000 (\$11,250 Academic year - \$3,750 for summer session) and the third year stipend will be \$16,000 (\$12,000 Academic year - \$4,000 summer session). Payment for tuition and fees will remain unchanged for second and third year.
6. In the event the Government should desire significant changes which would affect price, delivery schedule, or terms and provisions in the overall project's work content or scope, this subcontract may be amended appropriately in accordance with Section H, Provision Number 14 of this agreement.
7. (FFP Contract)
The contract price is \$17,500.00.
8. OPTION PROVISION: UES is hereby granted the right to obtain the performance of the work described in Section B, SUPPLIES, SERVICES and PRICES. UES may require the subcontractor to perform Options at the fixed price set forth in Section B by issuing a unilateral subcontract modification to the subcontractor on or before 1 September 1990 for Option I and 1 September 1991 for Option II.

9. LIMITATION OF UES OBLIGATION: Of the total subcontract price, the sum of \$17,500.00 is presently available for payment and is allotted to this subcontract. It is anticipated that from time to time additional funds will be allotted to this subcontract until the total price of this subcontract is allotted. It is contemplated that the funds presently allotted to this subcontract will cover the work to be performed through 31 August 1990.
10. Choice of Law: Irrespective of the place of performance, this Subcontract will be construed and interpreted according to the federal common law of government contracts enunciated and applied by federal judicial bodies, boards of contract appeals, and quasi-judicial agencies of the federal government. To the extent that the federal common law of government contracts is not dispositive, the laws of the state from which this subcontract is issued shall apply.
11. DISPUTES: Either party may litigate any dispute arising under or relating to this Subcontract before any court of competent jurisdiction. Pending resolution of any such dispute by settlement or by final judgment, the parties shall proceed diligently with performance. Subcontractor's performance shall be in accordance with UES's written instructions. All references to disputes procedures in Government clauses incorporated by reference shall be deemed to be superseded by this clause.
12. ENTIRE CONTRACT: This Subcontract and any documents specifically incorporated herein constitute the entire agreement between the parties. This written agreement supersedes all other prior agreements, written or oral, between the parties related to the subject matter hereof unless specifically otherwise provided herein. The terms and obligations created hereby shall be changed only in writing as a modification to this agreement, duly executed by the parties hereto.
13. ACKNOWLEDGEMENT OF SPONSORSHIP
 - a. The Subcontractor agrees that in the release of information relating to this contract a copy will be provided to UES and such release shall include a statement to the effect that the project or effort depicted was or is sponsored by the agency set forth below.

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
BOLLING AFB DC

b. For the purpose of this clause, "information" includes but is not limited to, news releases, articles, manuscripts, brochures, advertisement, still and motion pictures, speeches, trade association meetings, symposia, etc.

14. Amendments to this subcontract agreement may be negotiated between UES and Subcontractor. Such amendments must be in writing, and upon execution by both parties, will become a part of this subcontract agreement and will be subject to all other applicable terms and conditions of this subcontract agreement.

SECTION I
GENERAL PROVISIONS

1. Subcontractor agrees to perform the services in accordance with the following Federal Acquisition Regulations (FAR) clauses, which are incorporated herein by reference with the same force and affect as though herein set forth in full. The clauses applicable to this Subcontract Agreement are limited to only those which contain a specific subcontract/purchase order flow down requirement as may be contained in each particular clause. Except with respect to "Audit Negotiation" and "Examination of Records by Comptroller General," the Special Provisions and FAR clauses, including Special Clauses in full text, shall be deemed to be modified as appropriate to substitute the word "UES" for "Government" or "Contracting Officer," substitute the word "Subcontractor" for "Contractor," and substitute the word "Subcontract" for "Contract."

I Federal Acquisition Regulation Clauses

<u>FAR Paragraph No.</u>	<u>Clause Title</u>	<u>Date</u>
52.212-8	Defense Priority and Allocation Requirements	May 1986
52.215-1	Examination of Records by Comptroller General	April 1984
52.215-2	Audit-Negotiation	April 1984
52.222-4	Contract Work Hours and Safety Standards Act--Overtime Compensation	March 1986
52.222-26	Opportunity	E q u a l
52.222-35	for Special	Affirmative Action
52.222-36	Disabled and Vietnam Era Veterans	Affirmative Action
of Handicapped	Workers	

<u>FAR Paragraph No.</u>	<u>Clause Title</u>	<u>Date</u>
52.223-2	Clean Air and Water	April 1984
52.227-1	Authorization and Consent Alternate I (April 1984)	April 1984
52.227-2	Notice and Assistance Regarding Patent and Copyright Infringement	April 1984
52.227-11		Patent Rights -
Retention by the	April 1984	
	Contractor (Short Form)	
52.243-1	Changes-Fixed Price	April 1984

SECTION J
DOCUMENTS, EXHIBITS & ATTACHMENTS

<u>Attachment Number</u>	<u>Description</u>	<u>Date</u>
1.	Notification Letter	July 1989
2.	Progress Certification	Undated



Universal Energy Systems, Inc.

17 July 1989

Dear _____:

Congratulations on being selected to participate in the Department of Defense funded National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, under the sponsorship of the Air Force.

Universal Energy Systems, Inc. (UES) has been selected by the Air Force Office of Scientific Research (AFOSR) to administer the Academic Year of your fellowship. UES is currently in the process of preparing a subcontract with _____ for your tenure on the NDSEG fellowship. You will be kept informed as this progresses. We will establish the fellowship with _____ so that your stipend will be paid directly to you from the university. Your obligations under this program are to make satisfactory progress each year toward your PhD, provide input to UES at the completion of each academic year, and provide the Air Force with a copy of your thesis at the completion of your PhD.

I have enclosed a Fellowship Conditions Form, please sign and return this to UES at your earliest convenience.

I have also attached information on additional opportunities that AFOSR is offering to the NDSEG fellows. Please review this information and notify UES of your interest. Since an acceptance of the opportunity will increase your yearly stipend, it affects our negotiations with your university. Your assistance in notifying UES as soon as possible will greatly help us with the negotiations with your university.

If you have any questions regarding these matters, please do not hesitate to contact the undersigned, or Ms. Sue Espy (Program Administrator) at 1-800-533-7532 or in Ohio 513-426-6900.

Congratulations !!!!! and good luck on your continued efforts in graduate school.

Sincerely,

UNIVERSAL ENERGY SYSTEMS, INC

Rodney C. Darrah
Program Director

xc: Rod Darrah
UES contracts
Lt. Col. Claude Cavender
for the University Subcontract

NATIONAL DEFENSE SCIENCE AND ENGINEERING
GRADUATE FELLOWSHIP
CONDITIONS FORM

I understand that the following obligations and conditions concerning my Department of Defense National Defense Science and Engineering Graduate Fellowship.

1. Conditions of my fellowship for each fellowship year is contingent on my satisfactory progress toward a PhD in _____.
2. I will provide Universal Energy Systems, Inc. (UES) with a statement of progress each academic year.
3. I will report any significant results, awards, honors, etc. to UES.
4. In the event that I must withdraw from the fellowship program, I will provide UES with a statement of the reasons for my withdrawal and a statement of my future employment and educational plans.
5. I will provide the Air Force with a copy of my PhD thesis upon the completion of my PhD work.

Signature

Date

Certification Needed for Each Academic Term
CERTIFICATION OF ACADEMIC PROGRESS

Fellow: _____
University: _____

Semester/Academic Term: _____

Subcontract: S-789-DoD- _____

Fellow to complete

1. Courses - Give description of courses and grades received. (Attach sheet if extra space is needed.)

2. Give a description of research and progress toward dissertation. (Attach sheets if extra space is needed).

3. Give a brief statement of your involvement with the Laboratory _____ and _____. Also list any items of interest such as academic awards, publications, other information that can be used for NDSEG Fellowship Program.

"I certify that all information stated is correct and complete."

Signature/Fellowship Recipient

TYPED NAME/FELLOWSHIP RECIPIENT

CERTIFICATION OF ACADEMIC PROGRESS

"I certify that _____ is making satisfactory academic progress toward a Ph.D. in the area of _____ for the _____ 19__ academic term."

Signature/Advising Professor

TYPED NAME/TITLE OF ADVISING PROFESSOR



Universal Energy Systems, Inc.

11 October 1989

Student's Name & Address

Dear _____:

The laboratory and mentor assignment have been completed. You have been assigned to the _____. Information on your mentor is listed below:

MENTOR NAME

Mentor's Address

Mentor's Phone

You may contact your mentor about the program involvement of the laboratory. Your mentor may be contacting you as well.

As a reminder, you are eligible for a three day visit to the laboratory to discuss your research efforts with your mentor. If you wish to take advantage of this visit, arrangements must be made with UES. Contact Sue Espy at the UES office. All travel costs and a per diem allowance for the three days will be provided by UES. In addition, you are eligible to spend 10 to 14 weeks during the summers doing research at the laboratory. Your stipend will be paid to you by UES during the time that you are at the laboratory. All travel costs for the round trip to the lab for the summer work will be reimbursed by UES. For the time that you are at the lab, UES will provide an expense allowance. Arrangements for spending the summers at the laboratory need to be coordinated with your mentor.

If you have any questions concerning the mentor assignment, please do not hesitate to contact us as UES.

Sincerely,
UNIVERSAL ENERGY SYSTEMS, INC

Rodney C. Darrah
Program Director

xc: Chief Scientist
Mentor
Focal Point

r:\wp\ndseg\mentor.1tr